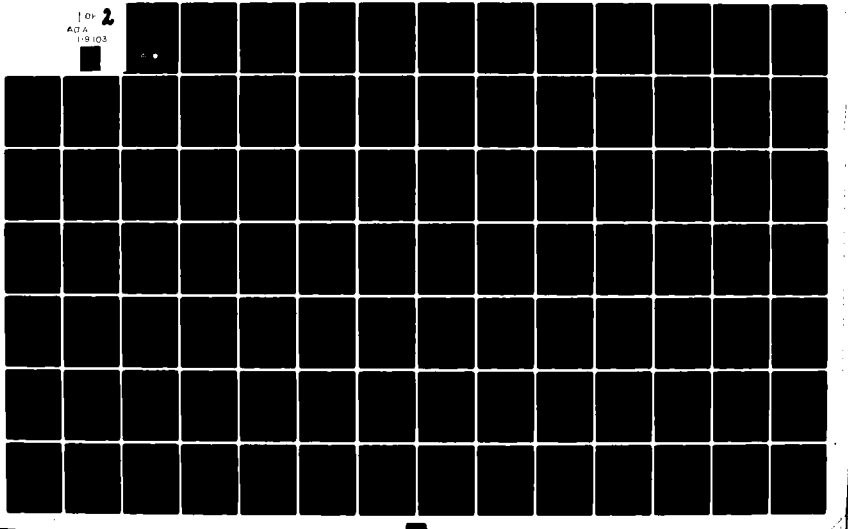
AD-A119 103 TEXAS UNIV AT AUSTIN APPLIED RESEARCH LABS F/G 12/1
PREPRINTS RESULTING FROM WORK UNDER SRO 106: NON-GAUSSIAN SIGNA--ETC(U)
AUG 82 P L BROKETT N00014-81-K-0145
UNCLASSIFIED ARL-TP-82-29 NL



(12)

PREPRINTS RESULTING FROM WORK UNDER SRO 106: NON-GAUSSIAN SIGNAL PROCESSING

TECHNICAL PAPERS UNDER CONTRACT N00014-81-K-0145

APPLIED RESEARCH LABORATORIES
THE UNIVERSITY OF TEXAS AT AUSTIN
POST OFFICE BOX 8029, AUSTIN, TEXAS 78712-8029

2 August 1982

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

Prepared for:

OFFICE OF NAVAL RESEARCH
DEPARTMENT OF THE NAVY
ARLINGTON, VA 22217

PREPRINTS RESULTING FROM WORK UNDER SRO 106: NON-GAUSSIAN SIGNAL PROCESSING

TECHNICAL PAPERS UNDER CONTRACT N00014-81-K-0145

APPLIED RESEARCH LABORATORIES
THE UNIVERSITY OF TEXAS AT AUSTIN
POST OFFICE BOX 8029, AUSTIN, TEXAS 78712-8029

2 August 1982

Prepared for:

OFFICE OF NAVAL RESEARCH
DEPARTMENT OF THE NAVY
ARLINGTON, VA 22217

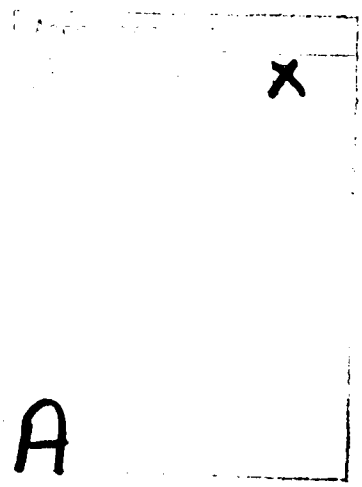


TABLE OF CONTENTS

ARL-TP-82-29
2 August 1982

THE LIKELIHOOD RATIO DETECTOR FOR NON-GAUSSIAN INFINITELY DIVISIBLE AND LINEAR STOCHASTIC PROCESSES

by

Patrick L. Brockett*

*Department of Finance
The University of Texas at Austin
Austin, Texas

This paper has been submitted to the
ANNALS OF STATISTICS for publication.

This research supported in part by the Office of Naval Research under Contract N00014-81-K-0145.

I. Introduction

The problem of optimum detection of signals in stochastic noise has been solved only in a relatively few cases. Most investigators assume the signal, or the noise, are Gaussian; however in many very important practical situations (e.g., radar, sonar or satellite transmission), this assumption does not hold (cf [10], [12], [25], [26]).

Lugannani and Thomas (1967) developed the class of linear processes, as a potential model for noise, and showed that the class was closed under linear transformations, a desirable property for modeling purposes. For a very specialized type of linear process, Eastwood and Lugannani (1977) were able to construct an approximation to the n-dimensional densities of two linear processes evaluated at $(t_1, t_2, \ldots, t_n)$. Consequently they were able to obtain a likelihood ratio test approximation for this special class of processes. It is of some considerable importance that the Middleton model for acoustical reverberation [18], [19], [20], [21] (which has been called the most complete theoretical model for reverberation [8]) is a linear process, as are several other models for noise derived from purely physical reasoning. Indeed, it was the physics of the acoustical noise process, and the goal of obtaining optimal signal processing which led to this problem.

In this paper we shall show how to identify a linear process as a subclass of infinitely divisible processes (i.e., processes with infinitely divisible finite dimensional marginal distributions). Using the results of Maruyama (1970), Brown (1971), Briggs (1975), Skorokhod (1964) and Veeh (1981), we are then able to explicitly calculate the Radon-Nikodym derivative of the measures determined on function space by two infinitely

divisible processes and to find its distribution. This extends the results of several authors and enables us to use the Neyman-Pearson lemma to obtain an optimal detector applicable directly to the sample paths of many stochastic process models derived in physics for noise processes.

II. Linear Processes

In this section we shall present a brief description of the linear process. A more extensive discussion and references are given in Lugannai and Thomas (1967).

The linear process $Y(t)$ is defined by the stochastic integral

$$(1) \qquad Y(t) = \int_a^b f(t,s)dX(s)$$

where $X(s)$ is a zero-mean, second order stochastically continuous process with independent increments. Additionally, $f(t,s)$ is real valued and square integrable with respect to $dV(s) = E|dX(s)|^2$. In brief, Y is an L_2 filtering of an independent increment process. We shall additionally assume f is L_2 continuous (so that $Y(t)$ is stochastically continuous). As particular cases we have the Gaussian process and the shot noise process.

Following the method used in Papoulis (1965) for shot noise process, one may determine the finite dimensional characteristic functions of the linear process (1) (cf Lugannani and Thomas (1967) or Eastwood and Lugannani (1977)). In the case without Gaussian component, they are given by

$$(2) \qquad \Phi_{\underset{\sim}{t}}(\underset{\sim}{u}) = \exp\{\int_a^b \int_0^\infty \{\exp(izw) - 1 - izw\}M(ds,dz)$$

where $\underset{\sim}{t} = (t_1,\ldots t_n)$, $\underset{\sim}{u} = (u_1,\ldots u_n)$ and $w = u_1f(t_1,s) + u_2f(t_2,s)+\ldots + u_nf(t_n,s)$. The measure M is the time-jump measure of the additive process X, i.e., $M((s_1,s_2] \times A)$ is the expected number of jumps (pulses) of the process X during the time interval $(s_1,s_2]$ with the magnitude (amplitude) in the Borel set A. See Gikhman and Skorokhod (1969) for a more detailed explanation of the Lévy measure M and its properties. We have deleted the Gaussian component of $Y(t)$ from (2) because it is independent of the non-Gaussian part, and it has already been treated extensively in the literature.

In the next section we shall describe the infinitely divisible processes and show that the linear process is, in fact, an infinitely divisible process of a very special type.

III. Infinitely Divisible Processes

The class of infinitely divisible stochastic processes was evidently first studied by Lee (1967), and subsequently studied by Maruyama (1970), Briggs (1975) and Veeh (1981). A stochastic process is called infinitely divisible if all of its n-dimensional marginal distributions are n-dimensional infinitely divisible random vectors. Gaussian processes are, of course, infinitely divisible and, for those second order processes without a Gaussian component, the following representation holds. By definition, for every finite subcollection $\lambda = \{t_1, t_2, \ldots, t_n\} \subseteq [a,b]$, there exists a random vector $\underset{\sim}{c}_\lambda$, and an n-dimensional Levy measure M_λ such that the characteristic function of $\left(Y(t_1), \ldots Y(t_n)\right)$ is

$$(3) \qquad \ln\Phi_\lambda(\underset{\sim}{u}) = i\underset{\sim}{u}'\underset{\sim}{c}_\lambda + \int\{\exp(i\underset{\sim}{u}'\underset{\sim}{x}) - 1 - iu'x\}dM_\lambda(\underset{\sim}{x}) \quad .$$

Here we have used a variant of the Kolmogorov representation valid for second order infinitely divisible vectors (cf Lukacs (1970) p. 119).

Let $\Lambda = \{\lambda = \{t_1, \ldots t_n\}\}$ denote the set of all finite subsets of [a,b]. The collection $\{(\underset{\sim}{c}_\lambda, M_\lambda), \lambda \in \Lambda\}$ uniquely determines the distribution of an infinitely divisible process Y, and vice versa (Maruyama (1970) Theorems 1 and 3). Using the partial ordering on Λ by inclusion, we obtain a system of projections $\{P_\lambda \ \lambda \in \Lambda\}$ from function space $\mathbb{R}^{[a,b]}$ onto the coordinate space $\mathbb{R}^\lambda$. The system of Levy measures $\{M_\lambda, \lambda \in \Lambda\}$ is consistent, and Maruyama shows that a measure Q may be defined on $\mathbb{R}^{[a,b]}$ as the projective limit of the collection $\{M_\lambda, \lambda \in \Lambda\}$. The σ-algebra on $\mathbb{R}^{[a,b]}$ is the usual product σ-algebra, and the construction of Q is similar to the usual construction in Kolmogorov's existence theorem for processes with given marginals.

Thus, corresponding to an infinitely divisible process there is a

function $c(t)$, and a measure Q living on function space such that for $\lambda = \{t_1,\ldots t_n\}$, $P_\lambda c = (c(t_1),\ldots c(t_n)) = \underset{\sim}{c}_\lambda$ and $QP_\lambda^{-1}(A) = M_\lambda(A)$ are the parameters of the infinitely divisible random vector $(Y(t_1),\ldots Y(t_n))$. We shall see that the projective limit Q is actually explicitly calculatable in the linear process case.

To obtain the appropriate representation of a linear process as an infinitely divisible process, we first manipulate the characteristic function given by (2) into the form of (3). Towards this end we first note that if $\lambda = \{t_1,\ldots,t_n\}$ is given, and M_λ is defined on $\mathbb{R}^\lambda$ via $M_\lambda(A)$ $= M(\{(s,z)\colon (zf(t_1,s),\ldots,zf(t_n,s)) \in A\})$. Then M_λ is a Levy measure on $\mathbb{R}^\lambda$ concentrated on the curve $(zf(t_1,s),\ldots,zf(t_n,s))$, $s \in [a,b]$, $z \in \mathbb{R}$. Moreover, the integral relationship $\int h(\underset{\sim}{x})dM_\lambda(\underset{\sim}{x}) = \int_a^b \int_{-\infty}^{\infty} h(zf(t_1,s),\ldots,zf(t_n,s))M(dz,ds)$ holds for measurable h. The fact that M_λ is indeed a Levy measure on $\mathbb{R}^\lambda$ follows (after some calculations) from the square integrability of f with respect to V and from the formula $\int h(s)dV(s) = \iint z^2 h(s)M(ds,dz)$ which relates the variance measure to the time-jump neasure M.

Now let us write $\underset{\sim}{f}(s) = P_\lambda f(\cdot,s) = (f(t_1,s),\ldots,f(t_n,s))$ and $\underset{\sim}{c}_\lambda = \underset{\sim}{0}$. We observe that

$$
\begin{aligned}
&i\underset{\sim}{u}'\underset{\sim}{c}_\lambda + \int_{\infty}^{\infty}\{\exp(i\underset{\sim}{u}'\underset{\sim}{x}) - 1 - i\underset{\sim}{u}'\underset{\sim}{x}\}\, M_\lambda(dx) \\
&= \iint\{\exp(iz\underset{\sim}{u}'\underset{\sim}{f}_\lambda(s)) - 1 - iz\underset{\sim}{u}'\underset{\sim}{f}_\lambda(s)\}M(ds,dz) \\
&= \iint\{\exp(izw) - 1 - izw\}M(ds,dz)
\end{aligned}
$$

where $w = u_1 f(t_1,s) + u_2 f(t_2,s) +\ldots+ u_n f(t_n,s)$ as before. Thus (2) is of the form (3). It follows that linear processes are in fact infinitely divisible processes. Moreover, we can determine the projective limit Q of the system of Levy measures $\{M_\lambda,\ \lambda \in \Lambda\}$. Namely, if $A \subseteq \mathbb{R}^{[a,b]}$ is a collection of functions, then $Q(A) = M(\{(s,z)\colon zf(\cdot,s) \in A\})$. This follows since if $B \subseteq \mathbb{R}^\lambda$, then $M_\lambda(B) = M(\{(s,z)\colon (zf(t_1,s),\ldots zf(t_n,s)) \in B\})$

$= M(\{(s,z): zP_\lambda f(\cdot,s) \in B\}) = Q(P_\lambda^{-1}(B))$, so that the λ^{th} coordinate projection of Q is M_λ.

In the next section we show how to use the results in Skorokhod (1964), Briggs (1975), Brockett, Hudson and Tucker (1978) and Brown (1978) to explicitly calculate the Radon-Nikodym derivation (likelihood ratio) of two linear processes on the function space $\mathbb{R}^{[a,b]}$. Note that since Y(t) is assumed to be stochastically continuous, the projective Levy measure Q on $\mathbb{R}^{[a,b]}$ is σ-finite by Proposition 3.1 in Maruyuma (1970). Along the way we extend the existing theorem on likelihood ratios for infinitely divisible processes.

IV. The Likelihood Ratio

We begin by sketching the construction of an infinitely divisible process Y as a limit of integrals of Poisson random measures. For details see Gikhman and Skorokhod (1969) or Maruyuma (1970).

Let π be a Poisson random measure on $\mathbb{R}^{[a,b]}$ which has the corresponding intensity measure Q, i.e., for any set $A \subseteq \mathbb{R}^{[a,b]}$ with $Q(A) < \infty$, $\pi(A)$ is a Poisson random variable with expectation Q(A). Moreover, if $A_1,\ldots,A_n$ are disjoint sets, then $\pi(A_1),\ldots,\pi(A_n)$ are independent random variables. See Kallenburg (1976) for details.

The random measure $\pi^*(A) = \pi(A) - Q(A)$ is used to give a pathwise representation of the second order linear process Y, namely we may write

$$(4) \qquad Y(t) = \lim_{\varepsilon \downarrow 0} \text{ in Prob.} \int_{A_\varepsilon} x(t)\pi^*(dx) \quad \text{where} \quad A_\varepsilon = \{x: |x(t)| \geq \varepsilon\} .$$

To see this, note that $\underset{\sim}{Y}_\lambda = (Y(t_1),\ldots,Y(t_n))$ has a characteristic function given by $\log \phi_{\underset{\sim}{Y}_\lambda}(\underset{\sim}{u}) = \log E(\exp i \sum_{j=1}^{n} u_j Y(t_j))$

$$= \int \{\exp(i \sum_{j=1}^{n} u_j X(t_j)) - 1 - i \sum_{j=1}^{n} u_j X(t_j)\} Q(dx)$$

$$= \int \{\exp(i\underset{\sim}{u}'P_\lambda x) - 1 - i\underset{\sim}{u}'P_\lambda x\} Q(dx)$$

$$= \int \{\exp(i\underset{\sim}{u}'y) - 1 - i\underset{\sim}{u}'y\} M_\lambda(dy).$$

Thus, the finite dimensional distributions given by the right hand side of (4) agree with those of the linear process.

We are now ready to calculate the likelihood ratio of two infinitely divisible processes without trend functions. The multivariate Lévy measures M_1 and M_2 induce (via projective limits) the measures Q_1 and Q_2 on function space as described earlier. The processes Y_i i = 1,2 determine measures

on function space via $\mu_i(A) = P[Y_i(\cdot) \in A]$, and we wish to determine when $\mu_1 << \mu_2$, and the corresponding density $\frac{d\mu_1}{d\mu_2}(x)$, $x \in \mathbb{R}^{[a,b]}$.

The following theorem generalizes the results of Briggs (1975) to include general infinitely divisible processes (not just those with non-atomic projective measures Q). As mentioned by Veeh (1981) it is not clear just how restrictive Briggs' assumptions are. Certainly for some sonar problems they are too restrictive. Additionally it generalizes the results of Veeh (1981) and Brockett, Hudson, and Tucker (1978). Moreover, by using the Maruyuma representation (4) for infinitely divisible processes, together with the results of Brown (1971) on Poisson point processes, we are able to substantially reduce the length and complexity of the proof of both the results of Briggs (1975), and of Brockett, Hudson and Tucker (1978). A key step is the following result (Lemma 1) for Poisson point processes due to M. Brown (1971).

First we recognize that a Poisson point process π on $(\mathcal{X},\mathcal{C})$ can be regarded as a probability measure on $(A,\mathcal{A})$ where A is the set of all countable subsets of atoms of $\mathcal{X}$ (multiple occurrence permitted) and $\mathcal{A}$ is the minimal σ-algebra containing all sets of the form $\{a \in A: \pi(C,a) = k\}$ $k = 0,1,\ldots,\infty$, $C \in \mathcal{C}$ and where $\pi(C,a)$, $a \in A$ is defined as the number of atoms comprising $a \in A$ which are contained in C (counting multiplicity). With this correspondence in mind, we have:

Lemma 1 (Brown 1971) Let π_1 and π_2 be Poisson point processes over $(\mathcal{X},\mathcal{C})$ with σ-finite mean measures Q_1 and Q_2. Then $\pi_1 << \pi_2$ if the following hold

i) $Q_1 << Q_2$

ii) $Q_1(B_t) < \infty$, $Q_2(B_t) < \infty$ for all $t > 0$

iii) $\int_{B_t^c} (\rho - 1)^2 dQ_2 < \infty$

where $\rho = \frac{dQ_1}{dQ_2}$ and $B_t = \{x: |\rho(x) - 1| > t\}$. Moreover, if Q_1 and Q_2 are finite measures, then $\frac{d\pi_1}{d\pi_2}(\{x_i\}) = \exp[-(Q_1(*) - Q_2(*)]\prod_{i=1}^{\pi_1(*)}\rho(x_i)$ where $\{x_i\} = \{x_i(\omega)\}$ is the particular realization under evaluation.

It should be noted that by the techniques Newman (1973) or Hudson and Tucker (1975) it can be shown that conditions (ii) and (iii) together are equivalent to the condition $\int(1 - \rho^{1/2})^2 dQ_2 < \infty$.

We now state our first results concerning the case with projective mean measures Q_1 and Q_2 finite. (This is the interesting case for acoustical signal detection using Middleton's model for surface reverberation for randomly rough surfaces, see Brockett and Wilson (1982).)

Theorem 1 (Non-Stationary Compound Poisson Case)

Suppose $Y_1(t)$ and $Y_2(t)$ are two stochastically continuous infinitely divisible processes with corresponding projective limit measures Q_1 and Q_2 finite.

a) If $Q_1 << Q_2$ and $\int x(Q_1 - Q_2)(dx) = 0$, then $\mu_1 = PY_1^{-1} << \mu_2 = PY_2^{-1}$. Moreover, using the representation (4),

$$\ell n \frac{d\mu_1}{d\mu_2}(Y_1(\cdot)) = \int \ell n\rho(x)\pi_1(dx) + Q_2(\mathbb{R}^{[a,b]}) - Q_1(\mathbb{R}^{[a,b]}) \text{ where } \rho(x) = \frac{dQ_1}{dQ_2}(x).$$

b) The μ_1 distribution of the log likelihood ratio in (a) is given via the characteristic function whose logarithm is

$\Psi(u) = iu[(Q_2 - Q_1)(\mathbb{R}^{[a,b]})] + \int(\exp\{iu\ell n\rho(x)\} - 1)\, Q_1(dx)$. Thus $\ell n \frac{d\mu_1}{d\mu_2}(Y_1(\cdot,\omega))$ is a translated compound Poisson random variable on $\mathbb{R}$ with intensity measure $\nu(A) = Q_1(\{x: \ell n\rho(x) \in A\})$.

Proof The proof of a) will follow immediately from Lemma 1 once we notice

that according to (4), and the corresponding representation for Poisson point processes as measures on a sequence space, we have $Y(\cdot,\omega) = S(\{x_i(\omega)\})$ where $S(\{x_i\}) = \sum_{i\geq 0} x_i - c_1 = \Sigma x_i - c_2$ and $\{x_i\}$ is a realization of the point process π. Here $c_1 = \int x dQ_1(x) = c_2 = \int x dQ_2(x)$ by the assumption $\int x d(Q_1 - Q_2)(x) = 0$. Now, under the assumptions of the theorem,

$$\mu_1 = \pi_1 S^{-1} << \mu_2 = \pi_2 S^{-1},$$

and by the lemma,

$$\frac{d\mu_1}{d\mu_2}(Y_1(\cdot)) = \frac{d\pi_1 S^{-1}}{d\pi_2 S^{-1}}(f) = \frac{d\pi_1}{d\pi_2}(s^{-1}f) = \frac{d\pi_1}{d\pi_2}(\{x_i(\omega)\})$$

$$= \exp[(Q_1(\mathbb{R}^{[a,b]}) - Q_2(\mathbb{R}^{[a,b]})] \prod_{\pi_1(\mathbb{R}^{[a,b]})} \rho(x_i(\omega)),$$

which is the formula in a) once the product is converted to integral form. Here we have used the fact that if ν and η are two measures on $*$ with $\nu << \eta$, and $S: * \to Y$, then $\nu s^{-1} << \eta s^{-1}$ and $\frac{d\nu s^{-1}}{d\eta s^{-1}}(y) = \frac{d\nu}{d\eta}(s^{-1}y)$. See Lemma 1 of Brockett, Hudson and Tucker (1978). Note that (ii) and (iii) of the lemma are obviously satisfied in this finite measure case.

To prove b) we simply notice that, according to the lemma, we are dealing with a Poisson sum (e.g., $\pi_1(\mathbb{R}^{[a,b]})$) of random variables, $\ell n\ \rho(x_i(\omega))$. The characteristic function now follows from routine calculations.

Using the results of Theorem 1, it is now just a short step to obtain the general theorem.

Theorem 2

Suppose $Y_1(t)$ and $Y_2(t)$ are two stochastically continuous infinitely

divisible processes with corresponding projective limit measures Q_1 and Q_2.

a) If i) $Q_1 << Q_2$ with $\rho(x) = \frac{dQ_1}{dQ_2}(x)$

ii) $\int x d(Q_1 - Q_2)(x) = 0$

iii) $\int (1 - \rho^{1/2}(x))^2 dQ_2(x) < \infty$

then $\mu_1 = PY_1^{-1} << \mu_2 = PY_2^{-1}$.

b) Under the conditions of a) $\ell n \frac{d\mu_1}{d\mu_2}(Y_1(\cdot)) = \int_{B_t^c} \ell n \rho(x) \pi_1^*(dx) + \int_{B_t} \ell n \rho(x) Q_1(dx)$

$+ \int_{B_t^c} [1 - \rho(x) + \ell n \rho(x)] Q_1(dx) + \int_{B_t} [1 - \rho(x)] Q_1(dx)$ where, as before,

$\pi_1^* = \pi_1 - Q_1$ and $B_t = \{x: |\rho(x) - 1| > t\}$.

c) The logarithm of the characteristic function of $\ell n (d\mu_1/d\mu_2)$ is

$$iu \int \left(1 - \rho(x) + \frac{\ell n \rho(x)}{1 + (\ell n \rho(x))^2}\right) dQ_1 + \int \left(e^{iuy} - 1 - \frac{iuy}{1+y^2}\right) Q_1 \circ g^{-1}(dy)$$

where $g = \ell n \rho$. Thus it is the translate of an infinitely divisible random variable with Levy measure $M(A) = Q_1(\{x: \ell n \rho(x) \in A\})$.

Proof Veeh (1981) proves a), or it could be derived from Lemma 1 in the previous manner. The proof of b) is given in Briggs (1975), and can follow from Theorem 1 by using her techniques. It should be noted that she does not explicitly state assumption (ii), although it is used in her proof. An alternative proof for b) can be constructed from Theorem 1 by using the techniques of Brockett, Hudson and Tucker (1978). The distribution in c) is obtained by a limiting argument from Theorem 1 after appropriately centering in a manner analogous to that used in Brockett, Hudson and Tucker (1978).

V. Linear Processes with Gaussian Components

In this section we show an alternative formula for the likelihood ratio when both driving functions X_1 and X_2 do have Gaussian components. Our development requires the additional assumption that the same filter is used on both processes, i.e., $f_1(t,s) = f_2(t,s)$. The key step in the development is the following very general result due to Skorokhod (1969, p. 245, Theorem 2) which of interest in its own right.

Lemma 2 (Skorokhod 1964) Suppose $X_1(t)$ and $X_2(t)$ are two stochastic processes inducing measures $\nu_1 = PX_1^{-1}$ and $\nu_2 = PX_2^{-1}$ on function space. Let S be a measurable mapping from function space to function space, and $Y_1 = SX_1$, $Y_2 = SX_2$ be two stochastic processes with induced measures $\mu_1 = PY_1^{-1}$ and $\mu_2 = PX_2^{-1}$. If $\nu_1 << \nu_2$ then $\mu_1 << \mu_2$ and $\frac{d\mu_1}{d\mu_2}(Y_2(t)) = E[\frac{d\nu_1}{d\nu_2}(X_2(t))|Y_2(t)]$.

We shall also use the results of Brockett, Hudson and Tucker (1978) and Brockett and Tucker (1977) for processes with independent increments, summarized in the Lemma 3. Let $X_1(t)$ and $X_2(t)$ be stochastically continuous processes with independent increments. Their characteristic functions are of the form

$$f_{X_1(t)}(u) = \exp\{iua(t) - \sigma^2(t)u^2/2 + \int(e^{iux} - 1 - \frac{iux}{1+x^2})M_t(dx)\}$$

(5) and

$$f_{X_2(t)}(u) = \exp\{iub(t) - r^2(t)u^2/2 + \int(e^{iux} - 1 - \frac{iux}{1+x^2}N_t(dx)\}.$$

The time-jump measures M and N are defined as before: (cf the sentence following equation (2)). The functions $\sigma^2(t)$ and $r^2(t)$ are continuous, non-negative, and nondecreasing and correspond to the Gaussian components $V_1(t)$ and $V_2(t)$ respectively.

Lemma 3 Suppose X_1 and X_2 are processes with independent increments over $[0,T]$, and $\nu_i = PX_i^{-1}$ $i = 1,2$ are the induced measures on function space $D[0,T]$.

a) $\nu_1 \sim \nu_2$ (i.e., $\nu_1 << \nu_2$ and $\nu_2 << \nu_1$) if and only if the following all hold:

(i) $M \sim N$

(ii) $\sigma^2(t) = \Gamma^2(t)$, $0 \leq t \leq T$,

(iii) $\int(1-(dM/dN)^{1/2})^2 dN < \infty$, and

(iv) for some $p \in L_2([0,T], d\sigma^2)$ and all $t \in [0,T)$,

$$a(t) - b(t) - \int \frac{x}{1+x^2}(M_t - N_t)(dx) = \int_0^t p(\theta)d\sigma^2(\theta).$$

b) If $\nu_1 \sim \nu_2$ on $D[0,T]$, then

$$\frac{d\nu_1}{d\nu_2} = \exp\{\int_0^T p(\theta)dV(\theta) - \frac{1}{2}\int_0^T p^2(\theta)d\sigma^2(\theta)$$

$$+ \sum_{n=1}^{\infty}(\Sigma\{\ell n \frac{dM}{dN}(s,J_s): \ 0 \leq s \leq T, \ J_s \in I_n \ - \int_{\Theta_n}(\frac{dM}{dN} - 1)dN)\},$$

$\Theta_n = [0,T] \times I_n$, $p(\theta)$ is as defined in a(ir), $I_n = [\varepsilon_n, \varepsilon_{n-1})U(-\varepsilon_{n-1}, -\varepsilon_n]$ with $\varepsilon_n \downarrow 0$, and J_s is the size of the jump at time s.

Using these results we may now state the conditions for equivalence directly in terms of the processes X_1 and X_2.

Theorem 3

Suppose Y_1 and Y_2 are linear processes given by (1) with the same filter function f for both Y_1 and Y_2. Let the distributions of the driving processes be given by (5). If (i) - (iv) of Lemma 3a all hold, then $\mu_1 = PY_1^{-1} \sim \mu_2 = PY_2^{-1}$ and $d\mu_1/d\mu_2$ is given by

$\frac{d\mu_1}{d\mu_2}(Y_1(t)) = E[\frac{d\nu_1}{d\nu_2}(X_1(t)) \mid Y_1(t) = \int f(t,s)dX_1(s)]$ where $\frac{d\nu_1}{d\nu_2}$ is found explicitly in Lemma 3b.

Proof By Lemma 3 it is clear that $\nu_1 = PX_1^{-1}$ & $\nu_2 = PX_2^{-1}$ and $\frac{d\nu_1}{d\nu_2}$ is as given.

We now use Lemma 2 to calculate the general likelihood ratio. Let $S: D[0,T] \to D[0,T]$ be defined by

$$(Sg)(t) = \int f(t,s)dg(s) = \lim \Sigma f(t,s_j)\{g(s_j) - g(s_j)\}$$

where the limit is as the mesh of the partition converges to zero. By the definition of Y_i given in equation (1), and by the results of Brockett and Hudson (1981) it is easily seen that S is defined a.s. with respect to both the measures $\nu_1 = PX_1^{-1}$ and $\nu_2 = PX_2^{-1}$. The stated likelihood ratio now follows from Lemma 2.

References

[1] Briggs, Vera Darlene, "Densities for Infinitely Divisible Random Processes," J. Mult. Anal. 5 178-205 (1975).

[2] Brockett, P. L., Hudson, W. N. and Tucker, H. G. ," The Distribution of the Likelihood Ratio for Additive Processes," J. Mult. Anal. 8 2 233-243 (1978).

[3] Brockett, P. L. and W. N. Hudson, "Variational Sums and Generalized Linear Processes," University of Texas Department of Finance Working Paper 81/82-2-36 (1981).

[4] Brockett, P. L. and Tucker, H. G., "A Conditional Dichotomy Theorem for Stochastic Processes with Independent Increments," J. Mult. Anal. 7 13-27 (1977).

[5] Brockett, P. L. and Wilson, G., "Likelihood Detection Using Middleton's Model of Surface Reverberation." In preparation. (1982).

[6] Brown, M., "Discrimination of Poisson Processes," Ann. Math Statist. 42 (1971).

[7] Eastwood, Lester F., Jr., and Lugannani, Robert, "Approximate Likelihood Ratio Detectors for Linear Processes," IEEE Trans. Inf. Th. IT-23, 482-489 (1977).

[8] Fortuin, L., "Survey of LIterature on Reflection and Scattering of Sound Waves at the Sea Surface," J. Acoust. Soc. Am. 47. 1209-1228 (1979).

[9] Gikhman, I. I. and Skorokhod, A. V., Introduction to the Theory of Random Processes, Saunders, Philadelphia (1969).

[10] Giordano, A. A. and Haber, F., "Modeling of Atmospheric Noise," Radio Science, 7, 1011-1023 (1972).

[11] Hudson, W. N. and Tucker, H. G., "Equivalence of Infinitely Divisible Distributions," Ann. Prob. 3, No. 1, 70-79 (1975).

[12] Kallenburg, Olav. Random Measures, Akademie-Verlag and Academic Press (1976).

[13] Kennedy, R. S., Fading Dispersive Communication Channels, Wiley, New York (1969).

[14] Lukacs, E., Characteristic Functions, Hafner Publ. New York (1970).

[15] Lee, P. M., Infinitely Divisible Stochastic Processes, Z. Wahrscheinlichkertstheorie Verw. Geb. 7, 147-160 (1967).

[16] Lugannani, R. and Thomas, J. B., "On a Class of Stochastic Processes which are closed under Linear Transformations," Information and Control, 10, 1-21 (1967).

[17] Maruyama, G., "Infinitely Divisible Processes," Th. Prob. Appl. XV No. 1-22 (1970).

[18] Middleton, D., "A Statistical Theory of Reverberation and Similar First Order Scattered Fields. Part I: Waveform an General Processes," IEEE Trans. Inf. Theory 13, 372-392 (1967).

[19] Middleton, D., "A Statistical Theory of Reverberation and Similar First Order Scattered Fields. Part II: Moments, Spectra, and Spatial Distributions," IEEE Trans. Inf. Theory 13, 393-414 (1976).

[20] Middleton, D., "A Statistical Theory of Reverberation and Similar First Order Scattered Fields. Part III: Waveforms and Fields," IEEE Trans. Inf. Theory 18, 35-67 (1972).

[21] Middleton, D., "A Statistical Theory of Reverberation and Similar First Order Scattered Fields. Part IV: Statistical Models," IEEE Trans. Inf. Theory 18, 68-90 (1972).

[22] Newman, C. M., "On the Orthogonality of Independent Increment Processes," in Topics in Probability Theory, Courant Institute of Mathematical Science (1973).

[23] Papoulis, A. Probability, Random Variables, and Stochastic Processes, McGraw Hill, New York (1965).

[24] Skorokhod, A. V., Random Processes with Independent Increments, Izdatel'stvo "Nauka," Moscow (1964) (in Russian).

[25] Trunk, G. V. and George, S. F., "Detection of Targets in Non-Gaussian Sea Clutter," IEEE Trans. Aerosp. Electron. Syst., AES-6, pp. 620-628 (1970).

[26] Van Trees, H. L., Detection, Estimation, and Modulation Theory, Part III: Radar-Sonar Signal Processing and Gaussian Signals in Noise, Wiley, New York (1971).

[27] Veeh, Jerry A., "Multidimensional Infinitely Divisible Measures," Ph.D. Dissertation, University of California, Irvine, Dept. of Mathematics (1981).

ARL-TP-82-30
2 August 1982

THE EQUIVALENCE OF WEAK, STRONG AND COMPLETE CONVERGENCE IN L1 FOR KERNEL DENSITY ESTIMATES

by

Luc Devroye*

*School of Computer Science
McGill University
Montreal, Canada

This paper has been submitted to the
ANNALS OF STATISTICS for publication.

This research supported by the Office
of Naval Research under Contract
N00014-81-K-0145.

THE EQUIVALENCE OF WEAK, STRONG AND COMPLETE CONVERGENCE IN L1 FOR KERNEL DENSITY ESTIMATES

by

Luc Devroye
McGill University

ABSTRACT.

Let f be a density on R^d, and let f_n be the kernel estimate of f,

$$f_n(x) = (nh^d)^{-1} \sum_{i=1}^{n} K((x-X_i)/h)$$

where $h=h_n$ is a sequence of positive numbers, and K is an absolutely integrable function with $\int K(x)dx = 1$. Let $J_n=\int|f_n(x)-f(x)|dx$. If $\lim_n h = 0$ and $\lim_n nh^d = \infty$, then $J_n \to 0$ completely as $n \to \infty$. If $J_n \to 0$ in probability as $n \to \infty$, And K is nonnegative and has compact support, then $\lim_n h = 0$ and $\lim_n nh^d = \infty$.

AMS 1970 Subject Classifications : Primary 60F15, Secondary 62G05.

Key Words and Phrases. Nonparametric density estimation, L1 convergence, kernel estimate, strong consistency.

1. INTRODUCTION.

The purpose of this paper is to point out that for the celebrated Parzen-Rosenblatt density estimate (Parzen, 1962; Rosenblatt, 1956) all types of L_1 consistency are equivalent. We consider a sample $X_1,\ldots,X_n$ of independent R^d-valued random vectors with common density f, and estimate f(x) by

$$f_n(x) = (nh^d)^{-1} \sum_{i=1}^{n} K((x-X_i)/h)$$

where $h=h_n$ is a sequence of positive numbers and K is a Borel measurable function satisfying $K \geq 0$, $\int K = 1$. The natural measure of the closeness of f_n to f is its L_1 distance,

$$J_n = \int |f_n(x)-f(x)|dx \ .$$

Our main result is

Theorem 1. Let K be a nonnegative Borel measurable function on R^d with $\int K(x)\,dx = 1$. Assume that K has compact support. Then the following conditions are equivalent :

(i) $J_n \to 0$ in probability as $n \to \infty$;

(ii) $J_n \to 0$ almost surely as $n \to \infty$;

(iii) $J_n \to 0$ completely as $n \to \infty$ (i.e. $\sum_n P(J_n > \varepsilon) < \infty$, all $\varepsilon > 0$)

(iv) $\lim_n h = 0$ and $\lim_n nh^d = \infty$.

Also, (iv) implies (iii) for all K that are absolutely integrable and satisfy $\int K(x)dx = 1$.

Remark 1. The condition that K has compact support is used in the proof of (i) => (iv). If K is any density on R^d, and (iv) holds, then we will see that $P(J_n > \varepsilon) = O(n^{-q})$ for all ε , $q > 0$.

Remark 2. If (i) holds for some K satisfying the conditions of Theorem 1, some f and some sequence h, then (iv),(iii),(ii) and (i) hold for all densities K and f.

We would like to clarify the connection between L_1 convergence and pointwise convergence almost everywhere. Pointwise convergence almost everywhere is a stronger concept as can be seen from the following variation of Scheffé's theorem (Scheffé, 1947) :

Lemma 1.(Glick, 1974). If f_n is a sequence of density estimates (i.e., for fixed n, f_n is a Borel measurable function of $x, X_1, \ldots, X_n$) and each f_n is a density in x, then $J_n \to 0$ in probability (almost surely) as $n \to \infty$ when $f_n(x) \to f(x)$ in probability (almost surely) for almost all x.

From Lemma 1 and well-known results about the pointwise convergence of kernel estimates, one can deduce L_1 convergence results that hold for all f (Devroye and Wagner, 1979). Unfortunately, for kernel estimates, strong pointwise convergence almost everywhere is a strictly stronger property than strong L_1 convergence, as can be deduced from Theorem 1 and Lemma 2 :

Lemma 2.(Deheuvels, 1974). If f is bounded and continuous, if K is bounded and Riemann integrable, if $h \to 0$ as $n \to \infty$ and $\inf_{n,k>0} (n+k)h_{n+k}^d/(nh_n^d) > 0$, then $f_n(x) \to f(x)$ almost surely for all x if and only if $nh_n^d/\log\log n \to \infty$ as $n \to \infty$.

Thus, the indirect type of proof based on Glick's Theorem (Lemma 1) does not give the sharpest possible L_1 convergence results except possibly in the weak case.

An analogue of Theorem 1 can be given for several other density estimates, such as recursive versions of the kernel estimate (surveyed, for example, in Devroye (1979)). The nearest neighbor estimates are not in L_1, and cannot be considered in the same light. In a series of papers, Abou-Jaoude (1976a, 1976b, 1976c) deals with a weak analogue of Theorem 1 for the histogram estimate. His main result is summarized below. Let Ψ_n be a sequence of countable partitions of R^d into λ-finite sets where λ is the Lebesgue measure : $\Psi_n=\{A_{n1},A_{n2},\ldots\}$, $0 < \lambda(A_{ni}) < \infty$. For $x \in R^d$, let $A_n(x)=A_{nj}$ if and only if $x \in A_{nj}$. Assume that the Borel sets are generated by $\bigcup_{m \geq n} \Psi_m$. Define the histogram estimate by

$$f_n(x) = (\lambda(A_n(x)))^{-1} \frac{1}{n} \sum_{i=1}^{n} I_{X_i \in A_n(x)} . \qquad (1)$$

Here I is the indicator function.

Theorem 2.(Abou-Jaoude, 1976a, 1976c). Let f be estimated by f_n, defined by (1). If $J_n \to 0$ in probability as $n \to \infty$ for all f, then

(i) for all Borel sets B of finite λ-measure, and all $\varepsilon > 0$, there exists n_0 such that for all $n \geq n_0$, there exists a set B_n in the σ-algebra generated by Ψ_n such that $\lambda(B \Delta B_n) < \varepsilon$;

(ii) $$\sup_{\substack{M > 0 \\ \text{all Borel sets } C \\ \text{with } \lambda(C)<\infty}} \limsup_n \left(\bigcup_{A_{nj}:\lambda(A_{nj})>M/n} A_{nj} \cap C \right) = 0.$$

If (i),(ii) hold, then $J_n \to 0$ completely as $n \to \infty$ for all f.

For example, if Ψ_n consists of all sets of the form $\prod_{i=1}^{d} [a_i b_n, (a_i+1)b_n)$ where $a_1,\ldots,a_d$ are all integers, and b_n is a sequence of positive numbers, then (i) and (ii) hold if and only if $b_n \to 0$ and $nb_n^d \to \infty$ as $n \to \infty$. We note here that the necessity of (i),(ii) in Theorem 2 follows from "$J_n \to 0$ in probability as $n \to \infty$ for all f". In contrast, the necessity of (iv) in Theorem 1 follows from "$J_n \to 0$ in probability as $n \to \infty$ for some f".

Abou-Jaoude's ground-breaking work excepted, very little is known about the L_1 properties of density estimates. Some information about the best possible rate of convergence to 0 of J_n can be found in Bretagnolle and Huber (1979). For more references on density estimation in general, the reader should consult Wegman (1972), Wertz (1978), Wertz and Schneider (1979) or Bean and Tsokos (1980).

2. PROOF OF THEOREM 1.

We will try to extract the key facts needed in the proof of Theorem 1. They are condensed in several Lemmas of independent interest. Lemmas 3 and 4 are integral and pointwise versions of the Lebesgue density theorem. Lemma 5 contains a crucial inequality for the multinomial distribution, and Lemma 6 establishes that (iv) => (iii) without the compactness condition for K. Lemma 7 is an L_1 version of the non-existence of unbiased kernel density estimates, and in Lemmas 8,9 and 10we present some unimportant rather basic inequalities. The necessity of (iv) for (i) is established in Lemma 11. Since (iii) => (ii) => (i), this would then complete the proof of Theorem 1.

Lemma 3.(L_1 version of Bochner's theorem).

Let K be an absolutely integrable function on R^d with $\int K(x)dx = 1$, and let $h=h_n$ be a sequence of positive numbers satisfying $\lim_n h = 0$. For each density f, we have

$$\lim_n \int |g_h(x)-f(x)|dx = 0$$

where

$$g_h(x) = h^{-d}\int K((x-y)/h)\, f(y)\, dy.$$

Proof of Lemma 3.

The proof is based on a technique of Kantorovich and Akilov (1964). I am grateful to Laszlo Győrfi for pointing this reference out to me. We let $C=\int |K(x)|dx$, and note that by a change of integral, for any function f,

$$\int |g_h(x)|dx \le \int\int h^{-d}|K((x-y)/h)||f(y)|dy\, dx = C\int |f(y)|dy.$$

For each $\epsilon > 0$ there exists a continuous function f vanishing outside* a compact set (say, S_{0R} where S_{xr} is the closed sphere of radius r centered at x) such that $\int|f(x)-f^*(x)|dx < \epsilon$. Thus, if we write $g_h(f,x)$ to make the dependence upon f explicit, then

$$\int|g_h(f,x)-f(x)|dx \le \int|g_h(f-f^*,x)|dx + \int|g_h(f^*,x)-f^*(x)|dx + \int|f^*(x)-f(x)|dx$$

$$\le (C+1)\int|f^*(x)-f(x)|dx + \int|g_n(f^*,x)-f^*(x)|dx$$

$$\le (C+1)\,\epsilon + \int|g_h(f^*,x)-f^*(x)|dx.$$

Thus, we need only show the Lemma for all functions f*. For each $\epsilon > 0$, find $\delta(\epsilon) > 0$ such that $||x-y|| < \delta(\epsilon)$ implies $|f^*(x)-f^*(y)| < \epsilon$. Thus, if f*=0 outside S_{0R}, then

$$\int|g_h(f^*,x)-f^*(x)|dx = \int\Big|\int_{\substack{||x||\le R\\ ||y||\le R}} h^{-d}K((x-y)/h)(f^*(y)-f^*(x))dy\Big|dx$$

$$\le \int_{||x||\le R}\Big(\int_{\substack{||y||\le R\\ ||x-y||\le\delta(\epsilon)}} + \int_{\substack{||y||\le R\\ ||x-y||\ge\delta(\epsilon)}}\Big|\,dx$$

$$\le \int_{||x||\le R}\Big(C\epsilon + C_1\int_{\substack{||y||\le R\\ ||x-y||\ge\delta(\epsilon)}} h^{-d}|K((x-y)/h)|dy\Big)\,dx$$

$$\le C\epsilon(2R)^d + C_1(2R)^d\int_{||hy||\ge\delta(\epsilon)}|K(y)|dy$$

$$= C\epsilon(2R)^d + o(1)$$

where $C_1 = \sup_x f^*(x)$. This concludes the proof of Lemma 3.

Lemma 4.(Lebesgue density theorem).

If f is a density on R^d and B is a compact set of R^d with $\lambda(B) > 0$, then

$$\lim_{h \downarrow 0} \lambda^{-1}(hB) \int_{x+hB} f(y)\, dy = f(x) \text{ , almost all } x.$$

Proof of Lemma 4.

We know that

$$\lim_{h \downarrow 0} \lambda^{-1}(S_{xh}) \int_{S_{xh}} |f(y)-f(x)|dy = 0$$

for almost all x, by the classical version of the Lebesgue density theorem (see for example, Stein (1970, pp. 62-63) or Wheeden and Zygmund (1977, pp. 100-109)). If S_{OR} is the smallest sphere containing B, then for almost all x,

$$\lambda^{-1}(x+hB) \int_{x+hB} |f(y)-f(x)|dy$$

$$\leq (\lambda(S_{OR})/\lambda(B))\, \lambda^{-1}(x+hS_{OR}) \int_{x+hS_{OR}} |f(y)-f(x)|dy$$

$$\to 0 \quad \text{as } h \downarrow 0.$$

Lemma 5.(A multinomial distribution inequality).

Let $(X_1,\ldots,X_k)$ be a multinomial$(n,p_1,\ldots,p_k)$ random vector. For all $\epsilon > 0$ and all k satisfying $k/n < (\epsilon/3)^2$, we have

$$P\left(\sum_{i=1}^{k} |X_i - E(X_i)| \geq n\epsilon\right) \leq \frac{1134}{n^2\epsilon^4}.$$

Proof of Lemma 5.

The proof is based upon a Poissonization. Let N be a Poisson (n) random variable independent of $U_1, U_2, \ldots$, a sequence of independent $\{1,\ldots,k\}$-valued variables distributed according to $P(U_1=i)=p_i$, $1 \le i \le k$. Let X_i be the number of occurrences of the value i among $U_1,\ldots,U_n$, and let X_i' be the number of occurrences of the value i among $U_1,\ldots,U_N$. It is clear that $X_1',\ldots,X_k'$ are independent Poisson random variables with means $np_1,\ldots,np_k$, and that $X_1,\ldots,X_k$ is a multinomial $(n,p_1,\ldots,p_k)$ random vector. Since $E(X_i) = np_i$, we have

$$\sum_{i=1}^{k} \frac{1}{n}|X_i-np_i| \le \sum_{i=1}^{k} \frac{1}{n}|X_i-X_i'| + \sum_{i=1}^{k} E(\frac{1}{n}|X_i'-np_i|)$$

$$+ \sum_{i=1}^{k} \frac{1}{n}(|X_i'-np_i|-E(|X_i'-np_i|))$$

$$\le \frac{1}{n}|N-n| + \sum_{i=1}^{k} \sqrt{p_i/n} + \sum_{i=1}^{k} \frac{1}{n}(|X_i'-np_i|-E(|X_i'-np_i|)). \quad (2)$$

By the Cauchy-Schwartz inequality, the middle term in (2) does not exceed $(k \sum_{i=1}^{k} p_i/n)^{1/2} = \sqrt{k/n} < \varepsilon/3$. Now, let $Z_i=\frac{1}{n}|X_i'-np_i|$, $Y_i=Z_i-E(Z_i)$. Then,

$$P(\sum_{i=1}^{k} \frac{1}{n}|X_i-np_i| \ge \varepsilon) \le P(\sum_{i=1}^{k} Y_i \ge \varepsilon/3) + P(|N-n| \ge n\varepsilon/3)$$

$$\le (3/\varepsilon)^4(\sum_{i=1}^{k} E(Y_i^4) + 6\sum_{i=1}^{k}\sum_{\substack{j=1 \\ j>i}}^{k} E(Y_i^2)E(Y_j^2) + n^{-4}E((N-n)^4)). \quad (3)$$

Now,

$$E((N-n)^4) = n+3n^2 \le 4n^2. \quad (4)$$

Also,

$$E(Y_i^4) = E(Z_i^4) - 4E(Z_i^3)E(Z_i) + 6E(Z_i^2)(E(Z_i))^2 - 3(E(Z_i))^4$$

$$\leq E(Z_i^4) + 6(E(Z_i^2))^2$$

$$= n^{-4}(np_i+3n^2p_i^2) + 6(np_i/n^2)^2$$

$$= 9p_i^2/n^2 + p_i/n^3 ,$$

and

$$E(Y_i^2) \leq E(Z_i^2) = p_i/n.$$

Combining these inequalities and resubstituting them into (3) shows that the left-hand-side of (3) is not greater than

$$(3/\varepsilon)^4 \left(4/n^2 + \sum_{i=1}^{k}(9p_i^2/n^2 + p_i/n^3) + 6\sum_{i=1}^{k}\sum_{\substack{j=1 \\ j>i}}^{k} p_ip_j/n^2\right)$$

$$\leq (3/\varepsilon)^4(4/n^2 + (9/n^2)(p_1+\ldots+p_k)^2 + 1/n^3)$$

$$\leq (3/\varepsilon)^4(14/n^2) = 1134/(n^2\varepsilon^4).$$

Lemma 6. For any density f on R^d, and any absolutely integrable function K with $\int K(x)dx = 1$, $J_n \to 0$ completely as $n \to \infty$ whenever

$$\lim_n h = 0 \quad \text{and} \quad \lim_n nh^d = \infty.$$

Proof of Lemma 6.

Let g_h be defined as in the statement of Lemma 3. By Lemma 3, it suffices to show that $\int|f_n(x)-g_h(x)|dx \to 0$ completely as $n \to \infty$. Let μ_n be the empirical probability measure for $X_1,\ldots,X_n$, and note that

$$f_n(x) = h^{-d}\int K((x-y)/h)\ \mu_n(dy).$$

For given $\varepsilon > 0$, find finite constants M,L,N,$a_1,\ldots,a_N$ and disjoint finite rectangles $A_1,\ldots,A_N$ in R^d such that the function

$$K^*(x) = \sum_{i=1}^{N} a_i\ I_{A_i}(x)$$

satisfies : $|K^*| \le M$, $K^*=0$ outside $[-L,L]^d$, and $\int|K(x)-K^*(x)|dx < \varepsilon$. Define g_h^* and f_n^* as g_h and f_n with K^* instead of K. Then

$$\int|f_n(x)-g_h(x)|dx \quad \le$$

$$\int|f_n(x)-f_n^*(x)|dx + \int|f_n^*(x)-g_h^*(x)|dx + \int|g_h^*(x)-g_h(x)|dx$$

$$\le \int h^{-d}\int|K^*((x-y)/h)-K((x-y)/h)|f(y)\ dy\ dx$$

$$+ \int h^{-d}\int|K^*((x-y)/h)-K((x-y)/h)|\mu_n(dy)\ dx + \int|f_n^*(x)-g_h^*(x)|dx$$

$$\leq 2\varepsilon + \int |f_n^*(x) - g_h^*(x)|\,dx$$

by a double change of integral. But if μ is the probability measure for f,

$$\int |f_n^*(x) - g_h^*(x)|\,dx \leq \sum_{i=1}^{N} |a_i| \int |h^{-d} \int_{x+hA_i} f(y)\,dy - h^{-d} \int_{x+hA_i} \mu_n(dy)|\,dx$$

$$\leq Mh^{-d} \sum_{i=1}^{N} \int |\mu(x+hA_i) - \mu_n(x+hA_i)|\,dx \ .$$

Lemma 6 follows if we can show that for all finite rectangles A of R^d,

$$h^{-d} \int |\mu(x+hA) - \mu_n(x+hA)|\,dx \to 0 \text{ completely as } n \to \infty.$$

Choose an A, and let $\varepsilon > 0$ be arbitrary. Consider the partition of R^d into sets B that are d-fold products of intervals of the form $[(i-1)h/N, ih/N)$, where i is an integer, and N is a fixed constant to be chosen later. Call the partition Ψ. If $A = \prod_{i=1}^{d} [x_i, x_i + a_i)$ and $\min_i a_i \geq 2/N$, then we let $A^* = \prod_{i=1}^{d} [x_i + 1/N, x_i + a_i - 1/N)$. Define

$$C_x = x+hA - \bigcup_{\substack{B \in \Psi \\ B \subseteq x+hA}} B \subseteq x + h(A - A^*) = C_x^* \ .$$

Clearly,

$$\int |\mu(x+hA) - \mu_n(x+hA)|\,dx$$

$$\leq \int \sum_{\substack{B \in \Psi \\ B \subseteq x+hA}} |\mu(B) - \mu_n(B)|\,dx + \int (\mu(C_x) + \mu_n(C_x))\,dx. \qquad (5)$$

The last term in (5) equals

$$2\lambda(h(A-A^*)) = 2h^d\lambda(A-A^*) = 2h^d\Big(\prod_{i=1}^{d} a_i - \prod_{i=1}^{d}(a_i-2/N)\Big)$$

$$= 2h^d\lambda(A)\Big(1-\prod_{i=1}^{d}(1-2/(Na_i))\Big) \le 4h^d\lambda(A)\sum_{i=1}^{d} a_i^{-1}/N \le \epsilon h^d$$

by choice of N. We used the fact that for any set C, and any probability measure ν on the Borel sets of R^d, $\int \nu(x+hC)\,dx = \lambda(hC)$. For any finite constant $R > 0$, we can rewrite the first term in (5) as

$$\sum_{\substack{B\in\Psi \\ B\cap S_{OR}\neq\phi}} |\mu_n(B)-\mu(B)| \int_{B\subseteq x+hA} dx + \int_{B\subseteq x+hA} dx\ (\mu_n(S_{OR}^c)-\mu(S_{OR}^c) + \mu(S_{OR}^c)). \tag{6}$$

Here $(.)^c$ denotes the complement of a set. Clearly, $h^{-d}\int_{B\subseteq x+hA} dx \le \lambda(A)$, and $\mu(S_{OR}^c) < \epsilon$ by our choice of R. Also,

$$P(\mu_n(S_{OR}^c)-\mu(S_{OR}^c) > \epsilon) \le e^{-2n\epsilon^2}$$

by Hoeffding's inequality for binomial random variables (Hoeffding, 1963). Finally, since the collection of sets $B\in\Psi$ with $B\cap S_{OR}\neq\phi$ has at most $(2RN/h+2)^d = o(n)$ elements, we see that by Lemma 5, for all n large enough,

$$P\Big(\sum_{\substack{B\in\Psi \\ B\cap S_{OR}\neq\phi}} |\mu_n(B)-\mu(B)| > \epsilon\Big) \le \frac{1134}{n^2\epsilon^4},$$

which is summable in n. Thus, for all $\epsilon > 0$, $P((6) > \epsilon)$ is summable in n. This concludes the proof of Lemma 6.

Remark 3. The bound $O(n^{-2})$ in Lemma 5 can be improved to $O(n^{-q})$ for any $q > 1$ by using a slightly more sophisticated technique. If this new bound is used for the first term of (6), we obtain the stronger result that under the conditions of Lemma 6,

$$\sum_{n=1}^{\infty} n^q P(J_n > \varepsilon) < \infty \quad , \text{ all } q, \varepsilon > 0.$$

Lemma 7. (Nonexistence of unbiased kernel density estimates).

Let K and f be arbitrary densities on R^d, and let g_h be defined as in Lemma 3. For any fixed $a > 0$,

$$\int |f(x)-g_a(x)|dx > 0.$$

Proof of Lemma 7.

It is obvious that $g_a(x)=E(f_n(x))$ is the density of X+Z where X has density f and Z is independent of X and has density $a^{-d}K(x/a)$. If ϕ and ψ are the characteristic functions of X and Z, then $\int |f(x)-g_a(x)|dx = 0$ implies $f=g_a$ for almost all x, and thus, $\phi(t)=\phi(t)\psi(t)$ for all t R^d. For $\phi(t)\neq 0$ (i.e., at least in a neighborhood of the origin), we must have $\psi(t)=1$. But then ψ cannot be the characteristic function of a density on R^d, and we have a contradiction.

Lemma 8. (A binomial distribution inequality).

If N is a binomial (n,p) random variable, then

$$E(|\frac{N}{n}-p|) \geq \begin{cases} e^{-4}\sqrt{p/(2\pi n)} & \text{when } \frac{1}{\sqrt{n}} \geq p \geq \frac{1}{n}, \ n \geq 4, \\ e^{-2}\, p & \text{when } p < \frac{1}{n}, \ n \geq 2. \end{cases}$$

Proof of Lemma 8.

Let $m = \underline{np}$, and note that

$$E(|\frac{N}{n} - p|) = \binom{n-1}{m} p^{m+1}(1-p)^{n-m} \qquad (7)$$

(Frame (1945); see also Johnson and Kotz (1969)). Assume first that $m \geq 1$. By Stirling's inequality, (7) is at least equal to

$$\frac{(n-m)^m}{m!} p^{m+1}(1-p)^{n-m} \geq \left(\frac{(n-m)pe}{m(1-p)}\right)^m \frac{p(1-p)^n}{\sqrt{2\pi m}} \; e^{-(12\,m)^{-1}}$$

$$\geq \left(\frac{npe}{m}\right)^m p(2\pi np)^{-1/2} \exp\left(-\frac{np}{1-p} - \frac{1}{12m}\right)$$

$$\geq (p/(2\pi n))^{1/2} \exp\left(m - \frac{np}{1-p} - \frac{1}{12m}\right) .$$

Now, $-1/(12m) \geq -1/12$, and $m - np/(1-p) \geq -1 - np^2/(1-p) \geq -1 - 1/(1-p) \geq -3$ when $p \leq 1/2$, which is guaranteed if $n \geq 4$. This proves the first inequality. When $m=0$, then

$$E(|\frac{N}{n} - p|) = p(1-p)^n \geq p(1-\frac{1}{n})^n \geq p \exp(-1/(1-\frac{1}{n})) \geq p\, e^{-2} \quad , \; n \geq 2.$$

Lemma 9.

If $X_1, \ldots, X_n$ are independent identically distributed random variables taking values in $\{0\} \cup [c, \infty)$ for some $c > 0$, then

$$E(|\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))|) \geq c\, E(|\frac{N}{n} - p|)$$

where N is a binomial (n,p) random variable and $p = P(X_1 \geq c)$.

Proof of Lemma 9.

Let X_1^* have the distribution of X_1 restricted to $X_1 \geq c$. Clearly, $E(X_1^*)=E(X_1 I_{X_1 \geq c})/p=E(X_1)/p$. If $N=\sum_{i=1}^{n} I_{X_i \geq c}$, then

$$E(|\frac{1}{n}\sum_{i=1}^{n}(X_i-E(X_i))|) = E(|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{N}{n}E(X_1^*) + (\frac{N}{n}p)E(X_1^*)|)$$

$$= E(|Y + (\frac{N}{n}p)E(X_1^*)|) \quad \text{(by definition of Y)}$$

$$\geq E(|E(Y|N) + (\frac{N}{n}p)E(X_1^*)|) = E(|\frac{N}{n}p|)E(X_1^*)$$

$$\geq c\, E(|\frac{N}{n}p|).$$

Lemma 10.

If Z is any random variable taking values on $[0,c]$, $c > 0$, then

$$\frac{E(Z)}{\varepsilon} \geq P(Z \geq \varepsilon) \geq \frac{E(Z)-\varepsilon}{c}, \text{ all } \varepsilon > 0.$$

Thus, if $Z_1, Z_2, \ldots$ is any sequence of $[0,c]$-valued random variables, then $Z_n \to 0$ in probability if and only if $E(Z_n) \to 0$.

Remark 4. The second statement of Lemma 10 is also a consequence of an inequality due to Feller (1971, pp. 152): $P(Z \geq \varepsilon E(Z)) \geq (1-\varepsilon)^2/E(Z^2) \geq (1-\varepsilon)^2/(cE(Z))$, $0 \leq \varepsilon \leq 1$.

Lemma 11. Let K and f be densities on R^d. If $J_n \to 0$ in probability as $n \to \infty$, then

$$\lim_n h = 0.$$

If also K has compact support, then

$$\lim_n nh^d = \infty.$$

Proof of Lemma 11.

Since $J_n \leq 2$ for all n, $J_n \to 0$ in probability if and only if $\lim_n E(J_n) = 0$ (Lemma 10). If g_h is defined as in Lemma 3, then

$$E(J_n) = E(\int |f_n(x)-f(x)|dx) \geq \int |E(f_n(x))-f(x)|dx = \int |g_h(x)-f(x)|dx.$$

Assume first that $\lim_n h = c > 0$. Clearly, when $c < \infty$,

$$\int |g_h(x)-f(x)|dx \geq \int |g_c(x)-f(x)|dx - \int |g_h(x)-g_c(x)|dx. \qquad (8)$$

As in the proof of Lemma 6, we find for each $\varepsilon > 0$ finite constants $M,L,N,a_1,\ldots,a_N > 0$ and disjoint finite rectangles $A_1,\ldots,A_N$ such that the function $K^* = \sum_{i=1}^{N} a_i I_{A_i}$ satisfies $|K^*| \leq M$, $K^*=0$ outside $[-L,L]^d$, and $\int |K^*(x)-K(x)|dx < \varepsilon$. If g_h^* is defined as g_h with K^* instead of K, we have

$$\int |g_h(x)-g_c(x)|dx \leq \int |g_h(x)-g_h^*(x)|dx + \int |g_c(x)-g_c^*(x)|dx$$

$$+ \int |g_h^*(x)-g_c^*(x)|dx$$

$$\leq 2\varepsilon + \int \Big| \sum_{i=1}^{N} a_i \mu(x+hA_i)h^{-d} - \sum_{i=1}^{N} a_i \mu(x+cA_i)c^{-d} \Big| \, dx$$

$$\leq 2\varepsilon + \sum_{i=1}^{N} h^{-d} a_i \int |\mu(x+hA_i)-\mu(x+cA_i)| dx$$

$$+ \sum_{i=1}^{N} c^{-d} a_i \int |\mu(x+hA_i)-\mu(x+cA_i)| dx + \int \sum_{i=1}^{N} a_i |c^{-d}-h^{-d}| \mu(x+cA_i) dx$$

$$\leq 2\varepsilon + \sum_{i=1}^{N} a_i (|c^{-d}-h^{-d}| \lambda(cA_i) + (h^{-d}+c^{-d}) \lambda(hA_i \Delta cA_i))$$

$$= 2\varepsilon + o(1). \tag{9}$$

Since $\varepsilon > 0$ was arbitrary, we conclude from (8) and (9) that

$$\liminf_n \int |g_h(x)-f(x)| dx = \int |g_c(x)-f(x)| dx. \tag{10}$$

But the right-hand-side of (10) is positive by Lemma 7. This contradicts the fact that the left-hand-side of (10) is 0. Therefore, no subsequence of h can converge to a finite positive constant. Assume next that $c = \infty$. We will show that

$$\liminf_n \int |g_h(x)-f(x)| dx \geq 1, \tag{11}$$

which again leads to a contradiction with $\lim_n E(J_n)=0$, so that no subsequence of h can diverge either. Therefore, $\lim_n h = 0$. To prove (11), we take an arbitrary $\varepsilon > 0$ and find M so large that $\int_{K(x)>M} K(x) dx < \varepsilon$. Define g_h^* and g_h^{**} as g_h after replacement of K by KI_A and KI_{AC} respectively, where $A=\{x: K(x) \leq M\}$. We have

$$\int|g_h(x)-f(x)|dx \geq \int|g_h^*(x)-f(x)|dx - \int g_h^{**}(x)\,dx$$

$$= \int|g_h^*(x)-f(x)|dx - \int_{K(x)>M} K(x)dx > \int|g_h^*(x)-f(x)|dx - \varepsilon.$$

Also, by Fatou's Lemma, since $g_h^*(x)=h^{-d}\int K((x-y)/h)I_A((x-y)/h)f(y)dy \leq Mh^{-d}\int_{x-hA} f(y)dy$ $\leq Mh^{-d} \to 0$ for all x,

$$\liminf_n \int|g_h^*(x)-f(x)|dx \geq \int \liminf_n |g_h^*(x)-f(x)|dx = 1,$$

so that (11) follows.

We now assume that K has compact support, that $\lim_n E(J_n) = 0$ and that $\lim_n nh^d = c$, $c \geq 0$. By Lemma 3, $\int|g_h(x)-f(x)|dx \to 0$ as $n \to \infty$. Thus,

$$\liminf_n E(J_n) \geq \liminf_n E(\int|f_n(x)-g_h(x)|dx). \tag{12}$$

We will show that the right-hand-side of (12) is positive for any $c \geq 0$, and thus, since the left-hand-side is 0, we must conclude that for all subsequences of h, $nh^d \to \infty$. Define $A=\{x: c\lambda(B)f(x) > 1\}$, where $B=\{x: K(x) \geq b\}$, and $b > 0$ is a constant that will be defined further on. Let $K^*=KI_B$, and g_h^* uses K^* in its definition, then

$$E(\int|f_n(x)-g_h(x)|dx) \geq E(\int|f_n^*(x)-g_h^*(x)|dx)$$

$$-E(\int h^{-d}\int|K((x-y)/h)-K^*((x-y)/h)|\mu_n(dy)\,dx)$$

$$-E(\int h^{-d}\int|K((x-y)/h)-K^*((x-y)/h)|f(y)dy\,dx)$$

$$= E(\int |f_n^*(x)-g_h^*(x)|dx) - 2\int |K(x)-K^*(x)|dx$$

$$= E(\int |f_n^*(x)-g_h^*(x)|dx) - 2\int_{K(x)<b} K(x)\,dx\ . \qquad (13)$$

By Lemma 9, $E(|f_n^*(x)-g_h^*(x)|) \geq bh^{-d}E(|\frac{N}{n}-p|)$ where N is binomial (n,p) and $p=P(K^*((x-X_1)/h) \geq b)=P(K((x-X_1)/h) \geq b)=P(X_1 \in x-hB)=\int_{x-hB} f(y)dy$. Assume that $n \geq 4$. Let $r=\min(e^{-2},e^{-4}/\sqrt{2\pi})$. By Lemma 8,

$$E(|f_n^*(x)-g_h^*(x)|) \geq \begin{cases} rbh^{-d}\sqrt{p/n} \geq rbh^{-d}/n\ , & 1/\sqrt{n} \geq p \geq 1/n, \\ rbh^{-d}p\ , & p < 1/n. \end{cases}$$

Thus,

$$E(\int |f_n^*(x)-g_h^*(x)|dx) \qquad (14)$$

$$\geq rbh^{-d}\left(\frac{1}{n}\lambda(x:n^{-1/2}>\int_{x-hB} f(y)dy>n^{-1}) + \int_{\int_{x-hB} f(y)dy<1/n} \int_{x-hB} f(y)dy\,dx\right).$$

Assume throughout that $\lambda(B) > 0$. Let L be the set of all x for which $\lambda^{-1}(x-hB)\int_{x-hB} f(y)dy$ tends to f(x) as $h \downarrow 0$. By the compactness of B, L includes almost all x (Lemma 4). We claim that the first term on the right-hand-side of (14) tends to $rb\lambda(A)/c$ when $c > 0$. Let C be a compact set of R^d, and let $A_n=\{x: 1/\sqrt{n} > \int_{x-hB} f(y)dy > 1/n\}$. Clearly, $A_n \cap L \to A \cap L$, and $nh^d \to c$. Thus, by the Lebesgue dominated convergence theorem, $\int_{A_n\cap L\cap C} dx \to \int_{A\cap L\cap C} dx = \int_{A\cap C} dx$, which is arbitrarily close to $\lambda(A)$ by choosing C large enough. For $c = 0$, we consider the second term of (14) only. First, for all $x \in L$ with $f(x) > 0$, the condition $\int_{x-hB} f(y)dy > \frac{1}{n}$ is violated for all n large enough.

Let $\overline{f}_h(x) = \int_{x-hB} f(y)dy$. The last term in (14) has the following limit infimum :

$$\liminf_n \; rb\lambda(B) \int_{\overline{f}_h(x)\leq 1/n} \lambda^{-1}(hB)\overline{f}_h(x)dx$$

$$\geq rb\lambda(B) \int \liminf_n I_{\overline{f}_h(x)\leq 1/n}\lambda^{-1}(hB)\, \overline{f}_h(x)\, dx$$

$$= rb\lambda(B) \int f(x)\, dx$$

$$= rb\lambda(B).$$

Here we used Fatou's Lemma and Lemma 4. Thus, the limit infimum of (13) is at least

$$rb\lambda(A)/c - 2 \int_{K(x)<b} K(x)dx \quad , c > 0,$$

$$rb\lambda(B) - 2 \int_{K(x)<b} K(x)dx \; , c = 0.$$

Choose $b > 0$ small enough such that $\lambda(B) > 0$ and that both expressions are strictly positive. This can be done since for $c > 0$,

$$\liminf_{b \downarrow 0} \lambda(A)/(\frac{1}{b}\int_{K(x)<b} K(x)dx) = \infty$$

and for $c = 0$,

$$\liminf_{b \downarrow 0} \lambda(B)/(\frac{1}{b}\int_{K(x)<b} K(x)dx) = \infty.$$

It is here that we use the compactness of K. This concludes the proof of Lemma 11.

3. DISCRIMINATION.

We would like to point out one important application of Theorem 1. In the discrimination problem, we are given a sequence $(X_1,Y_1),\ldots,(X_n,Y_n)$ of independent $R^d \times \{1,\ldots,M\}$-valued random vectors distributed as (X,Y) but independent of (X,Y). We construct an estimate Y from X and the data sequence, say, $Y = g_n(X)$. The probability of error for the given estimate and data sequence is $L_n = P(g_n(X) \neq Y | X_1,Y_1,\ldots,X_n,Y_n)$, and this is always at least equal to the Bayes probability of error $L^* = \inf_{g:R^d \to \{1,\ldots,M\}} P(g(X) \neq Y)$. If X has a density f, and if we construct the density estimates

$$f_{ni}(x) = (nh^d)^{-1} \sum_{j=1}^{n} K((x-X_j)/h)\, I_{Y_j=i} \ , \ 1 \leq i \leq M, \qquad (15)$$

and if we define $g_n(x)$ as the first integer i for which $f_{ni}(x) = \max_{1 \leq k \leq M} f_{nk}(x)$, then how is L_n related to L^* ? In other words, in what senses does L_n converge to L^* ? The simple rule mentioned here can be found under the name "potential function method" in the Russian literature (see e.g. Bashkirov, Braverman and Muchnik (1964)). Its properties were subsequently studied by Van Ryzin (1966), Rejtö and Révész (1973), Glick (1972,1976), Greblicki (1978), Devroye and Wagner (1980a,1980b) and Spiegelman and Sacks (1980). In this note, we can offer the following result :

Theorem 3.

Let K be an absolutely integrable function with positive integral over R^d, and let X have a density f. Then the discrimination rule defined by (15) satisfies

$$\sum_{n=1}^{\infty} n^q P(L_n - L^* > \varepsilon) < \infty \ , \ \text{all } q, \varepsilon > 0,$$

whenever

$$\lim_n h = 0, \text{ and } \lim_n nh^d = \infty.$$

Remark 5. Theorem 3 contains all previously known consistency results for the discrimination rule (15) that are based on the assumption that X has a density f. With additional conditions on K (i.e., $c_1 I_{S_{0,r_1}} \geq K \geq c_2 I_{S_{0,r_2}}$ for some $c_1, c_2, r_1, r_2 > 0$), we know that $L_n \to L^*$ in probability for all distributions of (X,Y) (Devroye and Wagner (1980); Spiegelman and Sacks (1980)). If we also ask that $r_1 = r_2$ and $nh^d/\log n \to \infty$, then $L_n \to L^*$ almost surely for all distributions of (X,Y). From our Theorem 3, it is clear that the condition $nh^d/\log n \to \infty$ is not needed whenever X has a density.

Proof of Theorem 3.

We introduce some new notation : $p_i = P(Y=i)$, $p_{ni} = \frac{1}{n}\sum_{j=1}^{n} I_{Y_j=i}$, f_i is the density of X given that Y=i, and $f_{n0} = \sum_{i=1}^{M} f_{ni}$. Then, by (12) of Devroye and Wagner (1980b),

$$L_n - L^* \leq \sum_{i=1}^{M} \int \left| \frac{f_{ni}(x)}{f_{n0}(x)} - \frac{p_i f_i(x)}{f(x)} \right| f(x)dx$$

$$\leq \sum_{i=1}^{M} \int |p_i f_i(x) - f_{ni}(x)|dx + \sum_{i=1}^{M} \int f_{ni}(x) \left| \frac{f(x)}{f_{n0}(x)} - 1 \right| dx$$

$$\leq \sum_{i=1}^{M} p_{ni} \int \left| f_i(x) - \frac{f_{ni}(x)}{p_{ni}} \right| dx + \int |f(x) - f_{n0}(x)|dx + \sum_{i=1}^{M} |p_i - p_{ni}|$$

$$\leq 2 \sum_{i=1}^{M} p_i \int \left| f_i(x) - \frac{f_{ni}(x)}{p_{ni}} \right| dx + \sum_{i=1}^{M} |p_i - p_{ni}|.$$

Let us look at i=1 only. By Hoeffding's inequality (Hoeffding, 1963), $P(|p_1 - p_{n1}| > \epsilon) \leq 2\exp(-2n\epsilon^2)$, all $\epsilon > 0$. Assume that $p_1 > 0$, and let $N = np_{n1}$.

Note next that $E(f_{n1}(x)/p_{n1}|N)=g_h(x)$ (defined in Lemma 3) if f is replaced by f_1. Thus,

$$\int|f_1(x)-\frac{f_{n1}(x)}{p_{n1}}|dx$$

$$\leq\int|f_1(x)-g_h(x)|dx\ I_{N>0} + \int|g_h(x)-\frac{f_{n1}(x)}{p_{n1}}|dx\ I_{N>0} + 2I_{N=0}. \tag{16}$$

The first term on the right-hand-side of the inequality tends to 0 as $h \to 0$ by Lemma 3. Conditional on N, the second term is distributed as $\int|E(f_N(x))-f_N(x)|dx$ where

$$f_N(x) = (Nh^d)^{-1} \sum_{i=1}^{N} K((x-X_i)/h)$$

and $X_1,\ldots,X_N$ are independent random vectors with common density f_1. In the proof of Theorem 1, we have seen that for every $\varepsilon > 0$ there exist positive constants c_i only depending upon ε, K and f_1 such that

$$P(\int|E(f_N(x))-f_N(x)|dx > \varepsilon|N) \leq c_1/N^q,$$

$$\text{valid when } (c_2/h+1)^d < c_3N.$$

Thus,

$$P(\int|g_h(x)-\frac{f_{n1}(x)}{p_{n1}}|dx\ I_{N>0} > \varepsilon) \leq P(N < np_1/2) + c_1(np_1/2)^{-q},$$

$$\text{valid when } (c_2/h+1)^d < np_1c_3.$$

Since $nh^d \to \infty$, the last inequality is valid for all n large enough. The term $P(N < np_1/2)$ does not exceed $\exp(-np_1^2/2)$ by Hoeffding's inequality, and the last term of (16) is treated similarly. Theorem 3 now follows by the arbitrariness of ε and q.

4. ACKNOWLEDGEMENTS.

This paper was written during the summer of 1981 while the author was visiting Applied Research Laboratories and the University of Texas at Austin. I wish to thank Charles Baker and the Office of Naval Research for the generous support.

REFERENCES

ABOU-JAOUDE, S.(1976a). Sur une condition nécessaire et suffisante de L_1-convergence presque complète de l'estimateur de la partition fixe pour une densité. C. R. Acad. Sci. Paris, Série A 283 1107-1110.

ABOU-JAOUDE, S.(1976b). Sur la convergence L_1 et L_∞ de l'estimateur de la partition aléatoire pour une densité. Ann. Inst. Henri Poincaré 12 299-317.

ABOU-JAOUDE, S.(1976c). Conditions nécessaires et suffisantes de convergence L_1 en probabilité de l'histogramme pour une densité. Ann. Inst. Henri Poincaré 12 213-231.

BASHKIROV, O.A., BRAVERMAN, E.M. and MUCHNIK, I.E.(1964). Potential function algorithms for pattern recognition learning machines. Automation and Remote Control 25 692-695.

BEAN, S.J. and TSOKOS, C.P.(1980). Developments in nonparametric density estimation. Int. Statist. Rev. 48 267-287.

BRETAGNOLLE, J. and HUBER, C.(1979). Estimation des densités : risque minimax. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 47 119-137.

DEHEUVELS, P.(1974). Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité. C. R. Acad. Sci. Paris, Série A 278 1217-1220.

DEVROYE, L.(1979). On the pointwise and the integral convergence of recursive kernel estimates of probability densities. Utilitas Mathematica 15 113-128.

DEVROYE, L. and WAGNER, T.J.(1979). The L1 convergence of kernel density estimates. Ann. Statist. 7 1136-1139.

DEVROYE, L. and WAGNER, T.J.(1980a). On the L1 convergence of kernel estimators of regression functions with applications in discrimination. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 51 15-25.

DEVROYE, L. and WAGNER, T.J.(1980b). Distribution-free consistency results in nonparametric discrimination and regression function estimation. Ann. Statist. 8 231-239.

FELLER, W.(1971). An Introduction To Probability Theory And Its Applications, vol.II, 2nd ed., John Wiley, New York.

FRAME, J.S.(1945). Mean deviation of the binomial distribution. American Mathematical Monthly 52 377-379.

GLICK, N.(1972). Sample-based classification procedures derived from density estimators. J. Amer. Statist. Assoc. 67 116-122.

GLICK, N.(1974). Consistency conditions for probability estimators and integrals of density estimators. Utilitas Mathematica 6 61-74.

GLICK, N.(1976). Sample-based classification procedures related to empiric distributions. IEEE Trans. on Information Theory IT-22 454-461.

GREBLICKI, W.(1978). Asymptotically optimal procedures with density estimates. IEEE Trans. on Information Theory IT-24 250-251.

HOEFFDING, W.(1963). Probability inequalities for the sum of bounded random variables. J. Amer. Statist. Assoc. 58 13-30.

JOHNSON, N.L. and KOTZ, S.(1969). Distributions In Statistics: Discrete Distributions. John Wiley, New York.

KANTOROVICH, L.V. and AKILOV, G.P.(1964). Functional Analysis In Normed Spaces. Pergamon Press.

PARZEN, E.(1962). On the estimation of a probability density function and the mode. Ann. Math. Statist. 33 1065-1076.

REJTŐ, L. and RÉVÉSZ, P.(1973). Density estimation and pattern classification. Problems of Control and Information Theory 2 67-80.

ROSENBLATT, M.(1956). Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27 832-837.

SCHEFFÉ, H.(1947). A useful convergence theorem for probability distributions. Ann. Math. Statist. 18 434-458.

SPIEGELMAN, C. and SACKS, J.(1980). Consistent window estimation in nonparametric regression. Ann. Statist. 8 240-246.

STEIN, E.M.(1970). Singular Integrals And Differentiability Properties Of Functions. Princeton University Press, Princeton, New Jersey.

VAN RYZIN, J.(1966). Bayes risk consistency of classification procedures using density estimation. Sankhya, Series A 28 161-170.

WEGMAN, E.J.(1972). Nonparametric probability density estimation: I. A summary of available methods. Technometrics 14 533-546.

WERTZ, W.(1978). Statistical Density Estimation. A Survey. Vandenhoeck and Ruprecht, Gottingen, Applied Statistics and Econometrics Series, Vol. 13.

WERTZ, W. and SCHNEIDER, B.(1979). Statistical density estimation : a bibliography. Int. Statist. Rev. 47 155-175.

WHEEDEN, R.L. and ZYGMUND, A.(1977). Measure And Integral. Marcel Dekker, New York.

Luc Devroye
School of Computer Science
McGill University
805 Sherbrooke Street West
Montreal, Canada H3A 2K6

ARL-TP-82-31
2 August 1982

ON THE INCONSISTENCY OF BAYESIAN NON-PARAMETRIC ESTIMATORS IN COMPETING RISKS/MULTIPLE DECREMENT MODELS

by

Barry C. Arnold*
Patrick L. Brockett**
William Torrez*
A. Larry Wright***

* Department of Statistics, University of California, Riverside, California
** Department of Finance and Applied Research Laboratories, The University of Texas at Austin, Austin, Texas
*** Department of Mathematics, University of Arizona, Tuscon, Arizona

This paper has been submitted to IEEE Transactions on Reliability Theory for publication.

This research supported in part by the Office of Naval Research under Contract N00014-81-K-0145.

§1. Introduction

The problem of analyzing the competing risks of various diseases has been studied in biometric, statistical and actuarial literature since Daniel Bernoulli first (1760) calculated the effect that the elimination of a specific disease (in this case smallpox) would have upon the general mortality rate structure of the population. This problem also is of vital importance in engineering reliability computations of complex systems and, under the guise of joint life status, in multiple decrement life tables in actuarial calculations (cf. Jordan (1975) Chapter 19).

The usual mathematical formulation of the problem is as follows. Given k possible mutually exclusive and collectively exhaustive risk factors (diseases, components which could cause system failure, etc.), let X_i, $i = 1, 2, \ldots, k$ denote the so-called net lives, i.e., the time which would expire until risk factor i operating alone would cause system failure (death of the individual or failure of a mechanical or electrical mechanism). The observable variable for each member of the population is $Z = \min_{1 \le i \le k} X_i$, the failure time of the system (the actual time of death of the person). Often (but not always) we are also able to ascertain which of the k competing risk factors is responsible for the failure. Using whatever data is available, our desire is to estimate the joint probability distribution measure for $X_1, X_2, \ldots, X_k$. Using this estimated probability measure we may consider interactions of the net lives, construct cause specific insurance policies, and subsequently set policy premiums for insuring losses, and design backup systems with appropriate reliability.

There are several problems in implementing the above procedure, the first and foremost of which is that in general the knowledge of the time and cause of death or failure alone does not usually uniquely determine the

joint measure. If we let J_ℓ denote the cause of failure of individual ℓ, and $Z_\ell = \min_{1 \le i \le k} X_i(\ell)$ the actual time of failure, then Tsiatis (1975) has shown that the data (Z_1,J_1), ..., (Z_m,J_m) need not uniquely determine the joint measure F_0 of $(X_1, \ldots, X_n)$. If the variables X_i are assumed to act independently, then Berman (1963) has shown that the data uniquely determines F_0. Within various parametric families, Nadas (1971) has shown this in the case where the X's are multivariate normal, and David and Moeschberger (1978) discuss other joint models such as multivariate exponential.

In many situations the assumption of mutual independent of the X's is not valid (morbidity and mortality from different causes tend to be related and component failures may tend to be related). When additional auxiliary information is available concerning the underlying multivariate dependence structure, it may not be compatible with the few uniquely identifiable distributions mentioned above. Moreover, in the case of vague information (e.g., the patient died but the cause was not known, and the time is only approximate) traditional methods fail to be able to incorporate this information. Additionally, one may not be precisely certain of the parametric class of distributions under study.

In such a situation one is tempted to adopt a Bayesian approach. Using prior information one would obtain a prior "best guess" for the joint probability law of the net lives, and, appealing to the traditional folklore, one would at least expect the posterior Bayesian estimator to be a consistent estimator of the true joint probability law. In this paper we shall derive the Bayesian non-parametric estimator using mixtures of Dirichlet processes (section 2). In section 3 we illustrate the technique on a bivariate censoring/death model with a joint Pareto law. In section 4 we show the (at first surprising) result that these Bayesian estimators are inconsistent if $n \geq 2$ except in possibly serendipitous cases.

§2. Dirichlet processes, mixtures of Dirichlet processes and Bayesian analysis of competing risks.

In order to do Bayesian analysis of competing risks, we must begin with a method for putting a prior distribution on the set of all joint probability measures on $(\mathbb{R}^+)^k$. Following Ferguson's (1973) lead we shall define such a prior distribution via the Dirichlet distribution. Since a stochastic process is, by definition, a function space valued random variable, we call the resulting prior distribution a Dirichlet process, and define it uniquely by specifying its finite dimensional marginals. We proceed as follows.

We will say that an r dimensional vector $X = (X_1, \ldots, X_r)$ has a Dirichlet distribution with parameter $\underset{\sim}{a} = (a_1, a_2, \ldots a_r) \in \mathbb{R}$ if the density of $(X_1, \ldots, X_{r-1})$ is of the form

$$f(x_i:\underset{\sim}{a}) = \begin{cases} \dfrac{\Gamma\left(\sum_{i=1}^{r} a_i\right)}{\prod_{i=1}^{r} \Gamma(a_i)} x_1^{a_1-1} x_2^{a_2-1} \cdots x_r^{a_r-1}, & \text{if } \sum_{i=1}^{r-1} x_i \leq 1, x_i > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

where $x_r = 1 - \sum_{i=1}^{r-1} x_i$.

We may define a Dirichlet stochastic process via its finite dimensional marginals as follows. Let α be a finite measure over $(\mathbb{R}^+)^k$, and let A_1, A_2, ..., A_m be an arbitrary partition of $(\mathbb{R}^+)^k$. The (random) measure F on $(\mathbb{R}^+)^k$ has a Dirichlet process distribution if the vector $(F(A_1), F(A_2), \ldots, F(A_m))$ has a Dirichlet distribution with a parameter $\underset{\sim}{a} = (\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_m))$ for every partition $A_1, \ldots, A_m$. Ferguson (1973) showed that this

collection of finite dimensional marginals uniquely defines a stochastic process $D(\alpha)$ called the Dirichlet process. Note that the mean is $E(D(\alpha)) = \alpha/\alpha((\mathbb{R}^+)^k)$ so that we may interpret α as our best prior assessment of the joint measure, and the total mass $\alpha(\mathbb{R}^{+k})$ as an indicator of our confidence in this prior estimate.

We apply Bayes Theorem using the prior Dirichlet process distribution on the parameter F, and the likelihood function of the data to obtain a posterior distribution on F which will be a mixture of Dirichlet processes, say $D(\beta)$, where β is a random measure. Our posterior estimate of the joint probability measure will be $m_p = E(D(\beta)) = E[\beta/\beta(\mathbb{R}^{+k})]$.

The following theorem can be derived from Antoniak (1974). It allows an iterative incorporation of data observations as they become available.

Theorem 1 If the prior distribution of the parameter F (the joint probability distribution measure) is a mixture $\int_\Lambda D(\alpha_\lambda)d\mu(\lambda)$ of Dirichlet processes with confidence coefficient $\alpha_\lambda(\mathbb{R}^{+k})$ constant in $\lambda\epsilon\Lambda$, an indexing set, then for any data of the form $(X_1, X_2, \ldots, X_k) \epsilon B$, where B is a Borel set with $\alpha_\lambda(B) > 0$ for all λ, the posterior distribution of F is also a mixture of Dirichlet processes,

$$\int_\Lambda \int_B D\left(\alpha_\lambda(\cdot) + I_{(\cdot)}(z)\right) \alpha_\lambda(dz|B)\mu(d\lambda) .$$

(Note that $\alpha_\lambda + I(z)$ is a random measure whose value on a set W is $\alpha_\lambda(W) + I_W(z)$.)

As a notational convenience we shall use the notation $D(\alpha)$ to denote a Dirichlet process with prior measure α or, in the case in which α is a random measure, to denote the mixture of Dirichlet processes. For example,

if we begin with non-random prior measure α and observe the data $\underset{\sim}{X}^{(1)} = (X_1, \ldots, X_k) \in B_1$, then the posterior measure has the distribution (by Theorem 1) given by the mixture

$$\int_B D(\alpha + I_{(\cdot)}(z))\, \alpha(dz|B_1)$$

$$= D(\alpha + I_{(\cdot)}(Z_1))$$

where Z_1 is a random variable with probability measure $\alpha(W|B_1) = \frac{\alpha(W \cap B_1)}{\alpha(B_1)}$. The mean or posterior measure is the expected value of the posterior process, m_p, and has the value on the set W of

$$m_p(W) = E[\alpha(W) + I_W(Z_1)]/(\alpha(\mathbb{R}^k) + 1)$$

$$= \frac{\alpha(W) + \alpha(W|B)}{\alpha(\mathbb{R}^k) + 1}$$

Similarly, the data $\underset{\sim}{X}^{(1)} \in B_1, \underset{\sim}{X}^{(2)} \in B_2$ yields a posterior distribution process which is the mixture (by Theorem 1 and the above calculation)

$$D(\alpha + I_{(\cdot)}(Z_1) + I_{(\cdot)}(Z_2|Z_1))$$

where Z_1 has the probability measure $\alpha(\cdot|B_1)$ as above, and $Z_2|Z_1 = z$ has the probability measure $\alpha_z(\cdot|B_2)$ where $\alpha_z = \alpha + I_{(\cdot)}(z)$. Using the fact that for a set W

$$E(I_W(Z_2|Z_1)) = E\left[\frac{\alpha(W|B_2) + I_{W \cap B_1 \cap B_2}(Z_1)}{\alpha(B_2) + I_{B_1 \cap B_2}(Z_1)}\right]$$

$$= \alpha(W|B_2)\alpha(B_2^c|B_1) + \alpha(W \cap B_2|B_1)\alpha^{(1)}(W|B_2)$$

$$+ \alpha^{(1)}(W|B_2)\alpha(B_2 \cap W^c|B_1) - \frac{\alpha(B_2 \cap W^c|B_1)}{\alpha^{(1)}(B_2)}$$

where $\alpha^{(1)}(W) = \alpha(W)+1$, it follows that the posterior measure satisfies

$$\begin{aligned} m_p(W)[\alpha(\mathbb{R}^{+k})+2] &= E[\alpha(W)+I_W(Z_1)+I_W(Z_2|Z_1)] \\ &= \alpha(W)+\alpha(W|B_1)+\alpha(W|B_2)\alpha(B_2^c|B_1) \\ &\quad + \alpha^{(1)}(W|B_2)\alpha(W\,B_2|B_1) + \alpha^{(1)}(W|B_2)\alpha(B_2\,W^c|B_1) \\ &\quad - \frac{\alpha(B_2\,W^c|B_1)}{\alpha^{(1)}(B_2)} \quad . \end{aligned}$$

In the case where $B_1 \cap B_2 = \phi$ this reduces to the simple relation

$$m_p(W) = [\alpha(W)+\alpha(W|B_1)+\alpha(W|B_2)]\Big/[\alpha(\mathbb{R}^{+k})+2].$$

Thus, analogously for data $\underset{\sim}{X}^{(1)} \in B_1$, $\underset{\sim}{X}^{(2)} \in B_2$, ..., $\underset{\sim}{X}^{(n)} \in B_n$ with mutually disjoint Borel sets B_1, ... B_n, the posterior Bayes estimate of the joint probability measure is

$$m_p(W) = [\alpha(W) + \sum_{i=1}^{n} \alpha(W|B_i)]/[\alpha(\mathbb{R}^{+k})+n] \quad . \tag{2.2}$$

As will be seen in section 3, the case in which $B_i \cap B_j = \phi$ occurs quite naturally in the context of competing risks.

§3. Analysis of survival time distribution with a random, non-independent censoring mechanism.

A particularly important special case of the preceding analysis involves the situation of only two factors -- occurrence of the event under study, or censoring (which includes loss to follow up, dropping out of the study, failure due to causes other than the event of interest, and even being alive at the termination of the study period). In general, censoring

is not independent of failure due to the cause of interest, and our previous theoretical development may be used to obtain an explicit formula for the survival curve in light of the prior information and the dependent censoring and failure time data.

For illustrative purposes, we shall assume the joint probability measure for the survival and censoring variables are related via a bivariate Pareto law. Thus, the joint measure of the set $[x_1,\infty) \times [x_2,\infty) = W$ is of the form

$$\alpha(W) = K[1+ (x_1/\sigma_1) + (x_2/\sigma_2)]^{-\gamma}, \quad x_1 > 0, \; x_2 > 0, \; \gamma > 0. \tag{3.1}$$

The constant $K = \alpha(\mathbb{R}^{+2})$, as before, has a "prior sample size" interpretation, or confidence interpretation. The constants σ_1 and σ_2 are scale parameters, and the parameter γ is interpreted as an association parameter. The bivariate Pareto law is a particularly beautiful model which assumes algebraically decreasing survival in each variable separately. If $\gamma > 2$ we may calculate (see, e.g., Arnold (1982) Chapter 6)

$$\begin{aligned} E(X_i) &= \gamma\sigma_i/(\gamma-1) \\ \mathrm{Var}(X_i) &= \gamma\sigma_i^2/[(\gamma-1)^2(\gamma-2)] \\ \mathrm{Cov}(X_1,X_2) &= \sigma_1\sigma_2/[(\gamma-1)^2(\gamma-2)] \\ \mathrm{Corr}(X_1,X_2) &= \gamma^{-1}. \end{aligned} \tag{3.2}$$

Additionally, the bivariate Pareto has the desirable property that the marginal and conditional distributions of X_1 and X_2 are univariate Pareto.

Let us now proceed to derive the posterior measure m_p for the joint survival censoring problem. The marginal posterior survival curve for the variable X_1 alone is then found by using the formulae

$$S(x) = m_p(X_1 \geq x)/m_p(\mathbb{R}^{+2}) \quad . \tag{3.3}$$

Returning for a moment to Theorem 1, we first note that if the data we obtain is of the form $\{X_1 > t, X_2 = t\}$ or $\{X_1 = t, X_2 > t\}$, then $\alpha(B) = 0$, so a slight modification is necessary. This is easily accomplished by taking $(X_1,X_2) \in B_\varepsilon$ where $B_\varepsilon = \{X_1 > t, |X_2 - t| \leq \varepsilon\}$ or $\{X_2 > t, |X_1 - t| \leq \varepsilon\}$ respectively, and then taking the limit as $\varepsilon \to 0$. The resulting equations for m_p now become

$$m_p(W) = [\alpha(W) + \sum_{i=1}^{n} \alpha(W|B_i)] \,/\, [\alpha(\mathbb{R}^{+2}) + n] \tag{3.4}$$

where $\alpha(W|B_i)$ is to be interpreted as the conditional α measure given the line segment $X_1 > t$, $X_2 = t$ or $X_2 > t$, $X_1 = t$ respectively, i.e.,

$$\alpha(W|B_i) = \lim_{\varepsilon \to 0} \frac{\alpha(W \cap B_\varepsilon)}{\alpha(B_\varepsilon)} \quad .$$

For the bivariate Pareto law (3.1) we have the conditional survival law of X_1 given $X_2 = x_2$ univariate Pareto with scale parameter $\sigma_1[1+x_2/\sigma_2]$ and exponent $\gamma + 1$, i.e.,

$$P[X_1 \geq x_1 | X_2 = x_2] = \left[1 + x_1 \Big/ \sigma_1[1+x_2/\sigma_2] \right]^{-\gamma-1} \quad . \tag{3.5}$$

If an individual is censored at time t, the information furnished is $X_1 > t$, $X_2 = t$, while if an individual fails the information furnished is of the form $X_1 = t$, $X_2 > t$. Using (3.5), we find the conditional probability $\alpha(X_1 \geq x | X_1 > t, X_2 = t) =$

$$\alpha P[X_1 > x | X_1 > t, X_2 = t] = \left[\frac{\sigma_1\sigma_2 + t\sigma_1 + t\sigma_2}{\sigma_1\sigma_2 + t\sigma_1 + x\sigma_2} \right]^{\gamma+1} \quad , \; x > t \quad .$$

We are now ready to specify the conditional survival function for X_1 given censored and failure time data.

Theorem 2 Consider the failure time-censoring model with bivariate prior random measure which follows a Dirichlet process with prior expectation measure $\frac{\alpha(\cdot)}{\alpha(\mathbb{R}^{+2})} = P_\alpha(\cdot)$ given by the Pareto law (3.1). Let $f_1, f_2, \ldots, f_r$ denote the observed failure times and $c_1, c_2, \ldots, c_m$ the observed censoring times for a sample of size r+m individuals, each of whom either fails, or is censored (but not both). Then the posterior random measure is a mixture of Dirichlet processes, and the posterior survival function for X_1 given the data and prior information is

$$S(x) = P[X_1 > x] = m_p([x,\infty)x\,\mathbb{R}^+) \Big/ m_{p(\mathbb{R}^{+2})}$$

$$= \frac{\alpha([x,\infty)x\,\mathbb{R}^+) + N_f(x) + \sum_{i=1}^{m} P_\alpha[X_1>x \mid X_1>c_i, X_2=c_i]}{\alpha(\mathbb{R}^{+2}) + r + m}$$

$$= \frac{\alpha(\mathbb{R}^{+2})[1+x/\sigma_1]^{-\gamma} + N_f(x) + \sum_{c_i<x} \left[\frac{\sigma_1\sigma_2+c_i\sigma_1+c_i\sigma_2}{\sigma_1\sigma_2+c_i\sigma_1+x\,\sigma_2}\right]^{\gamma+1} + N_c(x)}{\alpha(\mathbb{R}^{+2}) + r + m}$$

where $N_f(x)$ is the number of failures which occur after time x and $N_c(x)$ is the number of individuals censored after time x. The proof of Theorem 2 follows directly from Theorem 1, and from calculations involving the marginal and conditional distributions of multivariate Pareto laws.

It should be noted that vague end point information such as "death at time t, but cause of death unknown" can be written as $(X_1,X_2)\in B$ where $B = \{X_1=t, X_2>t\} \cup \{X_1>t, X_2=t\}$. Similarly if the exact time of death is only known to be $t_1 \le X_1 \le t_2$, we may incorporate this knowledge since it is also of the form $(X_1,X_2)\in B$ where $B = \{t_1 \le X_1 \le t_2, X_2 > t_2\}$. More exotic forms of information can also be incorporated by the same method.

§4. Inconsistency

It is tempting to assert that, as more and more data is collected, the influence of the assumed prior (determined by α) on the estimated value of F should diminish. Indeed, this is true for measures on $\mathbb{R}$ (cf. Susarla and Van Ryzin (1976) or Blum and Susarla (1977) or Bhattacharya (1981). Unfortunately, in the competing risks setting, such is rarely the case. We can illustrate this distressing situation in the context of our failure time-censoring model. Consider the posterior estimated survival function in Theorem 2. The expression

$$S(x) = P[X_1 > x] = m_p([x,\infty) \times \mathbb{R}^{+2}) \Big/ m_p(\mathbb{R}^{+2})$$

$$= \frac{\alpha([x,\infty) \times \mathbb{R}^{+}) + N_f(x) + \sum_{i=1}^{m} P_\alpha[X_1 > x | X_1 > c_i, X_2 = c_i]}{\alpha(\mathbb{R}^{+2}) + r + m}$$

is valid for a general prior measure α. It is not necessary that it correspond to a Pareto law. Imagine that we have available n observations (a random number r_n are failure times and the remaining random number m_n are censoring times). The strong law of large numbers applies and S(x) converges almost surely. However, after some calculation we find that the limit will coincide with the true marginal distribution F_1 if and only if true joint distribution and the prior guess α are related by

$$\int_0^x F(X_1 > v | X_2 = v) \frac{P_\alpha(X_1 > x | X_2 = v)}{P_\alpha(X_1 > v | X_2 = v)} \, d F_2(v)$$

$$= \int_0^x F(X_1 > x | X_2 = v) \, d F_2(v), \quad \text{all } x. \qquad (4.1)$$

A sufficient condition for this is that

$$\alpha = c F \quad . \tag{4.2}$$

A weaker sufficient condition is that

$$\frac{P_\alpha(X_1 > x | X_2 = v)}{P_\alpha(X_1 > v | X_2 = v)} = \frac{F(X_1 > x | X_2 = v)}{F(X_1 > v | X_2 = v)}, \qquad \text{all } x > v > 0. \tag{4.3}$$

We cannot expect to guess right, i.e., we cannot expect (4.2) to hold We can interpret (4.3) as the requirement that F and α assign proportional densities to half-lines of the form $\{(x,v): x > v\}$ where v is a fixed positive number. Again it would be excessive serendipity if this were to occur. Generally speaking the effect of α cannot be reasonably expected to "wash out" with increasing sample size.

The conclusion is quite logical in retrospect. It merely formalizes the fact that if our data is to give us no information about the conditional density on lines of the form $\{x > v\}$ then we will have no reasonable basis to change our prior beliefs about such densities. The implication of the preceding inconsistency result is that parametric analysis (either Bayesian or non-Bayesian) is mandated for the competing risk problem in system reliability. Much as we might like to proceed non-parametrically, even Bayesian analysis cannot give improved estimation results.

REFERENCES

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications cations to Bayesian nonparametric problems. Annals of Statistics 2, 1152-1174.

Arnold, Barry C. (1982). Pareto Distributions. Monograph to be published by International Co-operative Publishing House.

Berman, S. M. (1963). Notes on extreme values, competing risks, and semi-Markov processes. Annals of Math. Statist. 34, 1104-1106.

Bernoulli, D. (1760). Essai d'une nouvelle analyse de la mortalité causée par la vérole et des avantages de l'Inoculation pour la prevénir. Mem. de la Academie Royale de Science, 1-45.

Bhattacharya, P. K. (1981). Prior Distribution of a Dirichlet process from quantile response data. Annals of Statistics 9, 803-811.

Blum, J. and V. Susarla (1977). On the posterior distribution of a Dirichlet process given randomly right censored observations. Stoch. Processes Appl. 5, 207-211.

David, H. A. and M. S. Moeschberger (1978). The Theory of Competing Risks. Griffins Statistical Monographs, No. 39, London and High Wycombe.

Ferguson, T. S. (1973). A Bayesian analysis of some non-parametric problems. Annal Statist. 1, 209-230.

Jordan, C. W., Jr. (1975). Life Contingencies. Society of Actuaries, Chicago.

Nadas, A. (1971). The distribution of the identified minimum of a normal pair determines the distribution of the pair. Technometrics 13, 201-202.

Susarla, V. and J. Van Ryzin (1976). Non-parametric estimation from incomplete observations. J. Am. Stat. Assoc. 61, 897-902.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. Proc. Nat'l. Acad. Sci. 72, 20-22.

ARL-TP-82-32
2 August 1982

WHEN DOES THE βth PERCENTILE RESIDUAL LIFE FUNCTION DETERMINE THE DISTRIBUTION?

by

Barry C. Arnold*
Patrick L. Brockett**

* Department of Statistics, University of California, Riverside, California
** Department of Finance and Applied Research Laboratories, The University of Texas at Austin, Austin, Texas

This paper has been submitted to Operation Research for publication.

This research supported by the Office of Naval Research under Contract N00014-81-K-0145.

Abstract

Knowledge of the β^{th} percentile residual life function does not uniquely determine the underlying distribution. Additional information about the tail behavior of the distribution is needed. Knowledge of both the β_1 and β_2 percentile residual life functions does uniquely determine the distribution provided that β_1 and β_2 are non-commensurate. This corrects and extends the results of Schmittlein and Morrison (1981).

In reliability studies one often encounters situations in which analysis must be made concerning the lifetime of an entity or produced product. If the analysis must take place when the object under study has already aged t time units, then the problems regarding the non-failed units translate into inferences concerning the residual life at time t. Traditionally, this inference has involved the mean residual lifetime function $E(T-t|T\geq t)$, where T is the random variable denoting the total life of the object, and T-t is the residual life at time t. By a differential equation argument it is easily shown that the mean residual life uniquely determines the distribution of T. Thus, a linear mean residual lifetime uniquely characterizes the Pareto (II) distribution, a fact apparently first discovered by d'Addario (1939), and rediscovered periodically in the literature.

For some problems, it is convenient to work with the median or some other percentile of the residual life rather than the mean residual life. With the heavy tailed distributions commonly encountered in survival studies, a single long term survivor can have a marked effect upon the mean. Additionally, to calculate the mean life we need wait until every unit has failed, while the median residual life may be calculated as soon as the majority of the units have failed. For the β^{th} percentile of the residual life function we only need to wait until β percent of the units have failed.

In this note we shall show that in general the β^{th} percentile residual does <u>not</u> uniquely determine T. We shall, however, give conditions sufficient to ensure uniqueness does prevail. In particular, we shall give conditions which ensure that a linear median residual lifetime function is uniquely a Pareto (II) property. This will clarify the (incorrect) theorem in Schmittlein and Morrison (1981) where it is stated that the median residual life function is linear if and only if the distribution is Pareto (II).

<u>Non-uniqueness of the β^{th} percentile residual lifetime function</u>

The cumulative distribution function of the random variable T will be denoted by F, the survival or reliability function for T will be denoted by $\bar{F}(t) = P[T>t]$, and the β^{th} percentile of the residual life at time t will be denoted by $m_\beta(t)$. Throughout we shall assume F is continuous, strictly increasing and supported on $(0,\infty)$ to ensure that $m_\beta(t)$ is well defined for each β and t.

The relationship between $\bar{F}$ and $m_\beta(t)$ is

$$\bar{F}(m_\beta(t)+t) = (1-\beta)\bar{F}(t) \tag{1}$$

which is a special case of Schröder's functional equation $\psi(\phi(t)) = \delta\psi(t)$. This arises frequently in the theory of branching processes (see Kuczma (1968) Chapter VI, or Seneta (1968) for details regarding this equation).

Schmittlein and Morrison (1981) consider the equation (1) with $m_{\frac{1}{2}}(t) = a + bt$ and give a theorem stating that $m_{\frac{1}{2}}$ is linear if and only if T has a Pareto (II) distribution. This theorem is false. Since equivalents of the functional equation (1) arise so often in character·ation theorems, it is worthy of detailed study.

In considering the uniqueness properties of the β^{th} percentile residual life function, we note that two reliability functions $\bar{F}_1$ and $\bar{F}_2$ have residual function $m_\beta(t)$ if and only if

$$m_\beta(t) = \bar{F}_1^{-1}((1-\beta)\bar{F}_1(t)) - t = \bar{F}_2^{-1}((1-\beta)\bar{F}_2(t))-t, \tag{2}$$

or equivalently

$$\bar{F}_2(\bar{F}_1^{-1}((1-\beta)\bar{F}_1(t))) = (1-\beta)\bar{F}_2(t). \tag{3}$$

Thus uniqueness of the solution of (1) for a given residual life function $m_\beta(t)$ is equivalent to uniqueness in the functional equation (3) for $\bar{F}_2$ when $\bar{F}_1$ is given. Now letting $G(x) = \bar{F}_2(\bar{F}_1^{-1}(x))$, and $\bar{F}_1(t) = y$ in (3) yields the equivalent function equation

$$G((1-\beta)y) = (1-\beta)G(y) \quad 0 < y < 1. \tag{4}$$

Equation (4) arises frequently in characterizations of the exponential distribution (cf. Arnold (1971), Gupta (1973), Galambos and Kotz (1978)).

Next we note that given any monotone increasing continuous solution $\tilde{G}$ to (4) with $\lim_{y\to 0} \tilde{G}(y) = 0$, $\lim_{y\to 1} \tilde{G}(y) = 1$ we obtain a solution $\bar{F}_2$ to (3) by letting $\bar{F}_2(t) = \tilde{G}(\bar{F}_1(t))$, and this $\bar{F}_2$ also has $m_\beta(t)$ as its β-percentile residual life function. $\bar{F}_2$ is distinct from $\bar{F}_1$ provided $\tilde{G}(x) \neq x$. Arnold (1971) and J. S. Huang (1974) have noted that such solutions $\tilde{G}(x) \neq x$ can easily be contructed, even under the additional assumption that $\tilde{G}$ be differentiable on (0,1). Graphically,

a solution to (4) can be constructed by considering an arbitrary monotone function $h(y)$ for $(1-\beta) \leq y \leq 1$ which satisfies $h(1-\beta) = 1-\beta$, $h(1) = 1$, and then extending h via the functional equation (4) to successive subintervals $(1-\beta)^k \leq x \leq (1-\beta)^{k-1}$. Clearly then, there are a multitude of such solutions. An example is sketched in Figure 1. Galambos and Kotz (1978) give the equation for one such solution.

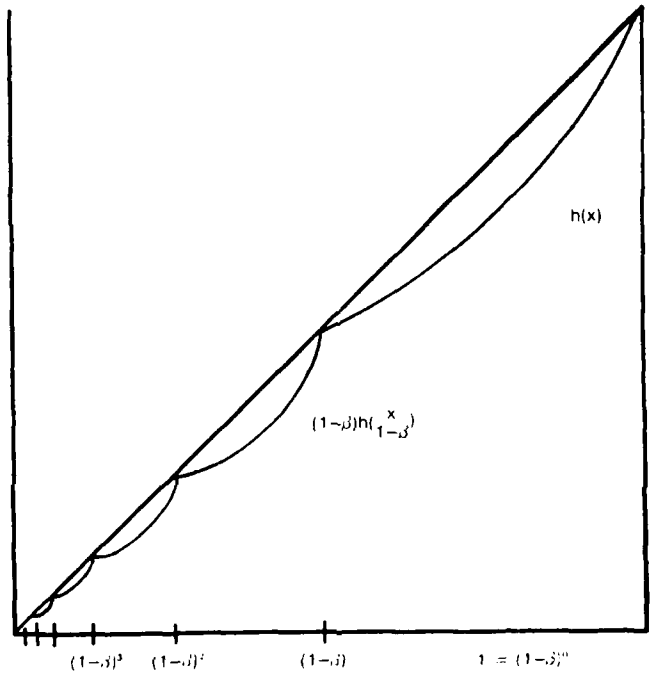


Figure 1 Construction of a nonlinear solution to (4)

Let G denote the class of all continuous increasing function G with $\lim_{x\to 0} G(x) = 0$, $\lim_{x\to 1} G(x) = 1$, with G satisfying (4). For each distribution $\bar{F}_1$, the entire class $F_1 = \{\bar{F}_2: \bar{F}_2 = G(\bar{F}_1), G \in G\}$ has the same β^{th} percentile residual life function $m_\beta(t)$. In particular, unlike the situation with the mean residual life function, the median residual (or more generally the β-th percentile) lifetime function <u>does not uniquely determine the distribution</u>! Thus, the stated theorem of Schmittlein and Morrison (1981) is incorrect, and the graphical method of Figure 1 can be used to construct many distributions other than the Pareto (II) distribution which possess linear median residual lifetime functions.

As a positive result we have the following theorem.

<u>Theorem 1</u>. Given two distributions $\bar{F}_1$ and $\bar{F}_2$

(a) If for some β, $\bar{F}_1$ and $\bar{F}_2$ both have β^{th} percentile residual life function $m_\beta(t)$, and also $\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = 1$, then $\bar{F}_1 = \bar{F}_2$.

(b) If $m_{\beta_1}(t)$ and $m_{\beta_2}(t)$ are the β_i^{th} percentile residual life functions of both $\bar{F}_1$ and $\bar{F}_2$ for any two noncommensurate percentiles β_1 and β_2, then $\bar{F}_1 = \bar{F}_2$.

Proof. If $\bar{F}_1$ is given, then the set of reliability functions $\bar{F}_2$ which solve (3) is the set F described previously. Conversely, given $\bar{F}_2$ we may construct a member $G\varepsilon G$ via $G(x) = \bar{F}_2(\bar{F}_1^{-1}(x))$. Thus uniqueness is equivalent to giving conditions on $\bar{F}_1$ and $\bar{F}_2$ which insure $G(x) = \bar{F}_2(\bar{F}_1^{-1}(x))\varepsilon G$ is the identity. A sufficient condition on G to insure this is $\lim_{x\to 0} G(x)/x = 1$ (for a proof see Arnold (1971) or Gupta (1973), however the plausibility of the result is immediate from figure 1). For $G(x) = \bar{F}_2(\bar{F}_1^{-1}(x))$ this condition is equivalent to $\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = 1$.

To prove (b), suppose that $\bar{F}_1$ and $\bar{F}_2$ have identical residual functions $m_{\beta_i}(t)$, i=1,2. Then $G(x) = \bar{F}_2(\bar{F}_1^{-1}(x))$ satisfies (4) for β_1 and β_2, the two non-commensurate values of β. Iterating (4) we find

$$G\left(\frac{(1-\beta_1)^k}{(1-\beta_2)^\ell}\, x\right) = \frac{(1-\beta_1)^k}{(1-\beta_2)^\ell}\, G(x) \text{ for all } \ell \text{ and } k.$$

Since the collection of ratios above is dense in [0,1] it follows that $G(x)=x$ for $x\varepsilon[0,1]$, and hence $\bar{F}_2 = \bar{F}_1$.

Theorem 2. Let $m_\beta(t)$ be given

a) For any fixed function g, let $F_g = \{\bar{F}: \lim_{x\to\infty} \frac{\bar{F}(x)}{g(x)} = c\}$. Then there exists at most one reliability function $\bar{F}\varepsilon F_g$ with β^{th} percentile residual life function $m_\beta(t)$.

b) Let $F_\alpha = \{\bar{F}: \lim_{x\to\infty} x^\alpha \bar{F}(x) = c \neq 0\}$. Then there exists at most one $\bar{F}\varepsilon F_\alpha$ with $m_\beta(t)$ as its β^{th} percentile residual life function.

Proof. Part (b) is just a particular case of (a) with $g(x) = x^{-\alpha}$, however it is stated separately because of its intimate connection with existence of moments. Namely, the class F_α is only slightly larger than the class of all distributions

with finite moments only up to α. It also contains the class implicitly assumed by Schmittlein and Morrison (1981) in their study of linear median residual life functions. The lacuna in their paper is between equations (10) and (11) in which they restrict their attention to distribution functions in a subset of F_α. By (b) there is uniqueness here.

For the proof of (a), assume $\bar{F}_1, \bar{F}_2 \in F_g$ are two solutions to (2). Then $G(x) = \bar{F}_2(\bar{F}_1^{-1}(x)) \in G$ solves (4), and $\lim_{x\to 0} G(x)/x = \lim_{y\to\infty} \frac{\bar{F}_2(y)}{\bar{F}_1(y)} = \lim_{y\to\infty} \frac{\bar{F}_2(y)/g(y)}{\bar{F}_1(y)/g(y)} = 1$. By the argument used in the proof of part (a) of Theorem 1, it follows that $\bar{F}_2 = \bar{F}_1$; i.e., there is at most one member of F_g with $m_\beta(t)$ as its residual life function.

To end on a happier note, we mention that the results of this paper give conditions under which the general residual lifetime function determines the distribution, and are not restricted to the linear median residual life and Pareto law. In particular the results apply to all the residual life functions outlined in Table 2 of Schmittlein and Morrison.

REFERENCES

Arnold, B. C. (1971). Two characterizations of the exponential distribution using order statistics, unpublished manuscript, Iowa State University, Ames, Iowa.

d'Addario, R. (1939). Un metodo per la reppresentazione analitica delle distribuzioni statistiche, Annali dell'Istituto di Statistica della Reale Universita de Bari 16, 3-56.

Galambos, J. and S. Kotz (1978). Characterizations of probability distributions, Springer-Verlag lecture notes in mathematics, Vol. 675, Berlin.

Gupta, R. C. (1973). A characteristic property of the exponential distribution, Sankhya B, 35, 365-366.

Kuczma, M. (1968). Functional equations in a single variable, Polish Scientific Publishers, Warsaw.

Schmittlein, D. C. and D. G. Morrison (1981). The median residual lifetime: A Characterization Theorem and an Application, Oper. Res., Vol. 29, pp. 392-399.

Seneta, E. (1968). Functional equations and the Galton-Watson process, Adv. Appl. Prob., Vol. 1, 1-42.

ARL-TP-82-33
2 August 1982

VARIATIONAL SUMS AND GENERALIZED LINEAR PROCESSES

by

Patrick L. Brockett*
William N. Hudson**

* Department of Finance and Applied Research Laboratories, The University of Texas at Austin, Austin, Texas
** Department of Mathematics, Auburn University, Auburn, Alabama

This paper has been submitted to Stochastics for publication.

This research supported by the Office of Naval Research under Contract N00014-81-K-0145.

1. Introduction

In this paper we consider properties of the stochastic process $Y(t) = \Sigma f(t,s,Js)$ where $Js = X(s) - X(s-0)$ is the (random) size of the jump, s is the (random) time of the jump of an additive stochastically continuous process X(s) with no Gaussian component, and f is an appropriate function. In the case $f(t,s,x) = xI_{[s \leq t]}$ (where I_A denotes the indicator function of the set A), we recapture the driving process X(t). The case f(t,s,x) = g(x) yields a variational sum for the process with independent increments. Such variational sums have been studied extensively in the literature under a variety of conditions on g and X(s) (e.g., [1], [2], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [17], [23], [24]). Analysis of Y yields information concerning the sample path properties of the processes X(t).

The choice $f(t,s,x) = h(t,s)x$ yields the representation $Y(t) = \int h(t,s)dX(s)$. Such processes are called linear processes by Eastwood and Lugannani (1977) and are useful in signal detection models. Another important applied motivation for considering general variational sum Y(t) outlined above is that the likelihood ratio of certain jump processes is of this form. Specifically, if $X_1(t)$ and $X_2(t)$ are two stochastically continuous processes with independent increments and "time-jump" measures M and N respectively (see section 2 for the definitions) then Brockett, Hudson and Tucker (1978) show that the likelihood ratio involves the variational sum

$$\sum_{n=1}^{\infty} \left(\Sigma \left\{ \ln \frac{dM}{dN}(s,Js): \ 0 \leq s \leq T, \ Js \in I_n \right\} - \int_{I_n} \left(\frac{dM}{dN} - 1 \right) dN \right)$$

which is of the above form with $f(t,s,x) = \ln \frac{dM}{dN}(s,x)$. Thus if a Neyman-Pearson detector is to be implemented, such variational sums need to be investigated.

An additional application of such generalized linear processes is the extension of Middleton's model for reverberation in underwater acoustical environments to include random amplitude components (cf. Middleton 1967a,b, 1972a,b). The derived model is precisely of this form.

Approximations to the variational sums using observed data (e.g., in order to implement the likelihood ratio detector for additive processes) necessitates investigation of the approximating sums

$$V_t(\Pi,f,X) = \sum_j f(t,s_j,\Delta X_j)$$

where $\Pi = \{s_0 < s_1 < s_2 \ldots < s_n\}$ is a partition of $[0,T]$ and $\Delta X_j = X(s_{j+1}) - X(s_j)$ is the increment of the process over the j^{th} subinterval. We shall investigate when $V_t(\Pi,f,x)$ converges a.s. to $Y(t) = \Sigma\, f(t,s,Js)$ as the norm of the mesh $||\Pi||$ goes to zero. This will also generalize the theorem in Riedel (1980) concerning existence of the stochastic integral as a limit of Stieltjes sums.

Most previous investigators of variational sums considered functions of the form $f(t,s,x) = |x|^p$ and processes $X(s)$ having stationary independent increments. Additionally most assumed a nested sequence of partitions $\Pi_n \to 0$. Our results generalize their results in that:

(i) our functions f allow dependence on the time as well as the size of the jump (this is necessary for the likelihood ratio problem in the non-stationary case),

(ii) we consider non-stationary driving processes $X(s)$,

(iii) we consider limits along a net of partitions $\{\Pi\} \to 0$ rather than nested sequences of partitions.

In section 2 we consider existence properties of the variational sums $\sum_s f(t,s,Js)$ in terms of f and the time-jump measure of the driving process X, and in section 3 we consider limits for the approximating sums $\sum_\Pi f(t,s_j,\Delta X_j)$.

2. Variational Path Properties

Throughout this paper we shall let X(t) be a separable stochastically continuous stochastic process with (not necessarily stationary) independent increments over [0,T], and without a Gaussian component. Without loss of generality we can and do assume $X(\cdot,w) \in D[0,T]$, i.e., the sample functions of X are right continuous with left limits at each point. We also assume X(0) = 0. Then X(t) is infinitely divisible with characteristic function

$$\Phi_{X(t)}(u) = \exp\{ia(t)u + \int_{(-\infty,\infty)} (e^{iux} - 1 - \frac{iux}{1+x^2})dM_t(x)\} \quad .$$

The parameters a(t) and M_t determine the distribution of X(t).

The time-jump measure of the process X is defined over $[0,T] \times R$ by $M((t_1,t_2] \times A) = M_{t_2}(A) - M_{t_1}(A)$. It is known that $M((t_1,t_2] \times A)$ is the expected number of jumps of X(s) for $s \in (t_1,t_2]$ with jump sizes $Js = X(s) - X(s-0) \in A$.

Our first results concern existence of Y(t), i.e., properties of f and M which will yield convergence of the sum defining Y(t). Theorem 1 below improves results due to Blumenthal and Getoor (1961) in the case $f(t,s,x) = |x|^\gamma$, and Fristedt (1967) in the case f(t,s,x) = h(x) where h is strictly increasing non-negative with h^{-1} convex. As a particular case it gives necessary and sufficient conditions for existence of a linear process much weaker than those in Lugannani and Thomas (1967).

Theorem 1 Let f be a measurable real valued function.

Then: a) $\sum_{s \in [0,T]} |f(t,s,Js)| < \infty$ a.s.

or $\sum_{s \in [0,T]} |f(t,s,Js)| = \infty$ a.s.

b) $\sum_{s\in[0,T]} |f(t,s,Js)| < \infty$ a.s.

if and only if

$$\int_{[0,T]\times[-1,1]} \{|f(t,s,x)| \wedge 1\}\, M(ds,dx) < \infty$$

(here $f \wedge 1$ denotes $\min\{f,1\}$).

Proof Without loss of generality, we may assume $f \geq 0$. Moreover, since there are only a finite number of jumps of absolute magnitude exceeding one, it follows that $\sum_s f(t,s,Js) < \infty$ if and only if $\sum_s \{f(t,s,Js): |Js| < 1\} < \infty$. Also, if we take $B_o = \{(s,x): f(t,s,x) > 1\}$, then $\sum\{f(t,s,Js): (s,Js) \in B_o\} < \infty$ if and only if $M(B_o) < \infty$. Thus, without loss we assume $0 \leq f(t,s,x) \leq 1$, and $|Js| \leq 1$ a.s. for the remainder of the proof.

Now, for $k = 1,2,\ldots$, let $B_k = \{(s,x): 2^{-k} < f(t,s,x) \leq 2^{-k+1}\}$. The random variables $N_k = \sum \{Js: (s,Js) \in B_k\}$ are independent Poisson variables with expectation $M(B_k)$.

The proof of a) follows from the Kolmogorov zero one law by writing $\sum_s f(t,s,Js) = \sum_{k=1}^{\infty}(\sum \{f(t,s,Js): (s,Js) \in B_k\})$ as a sum of independent random variables. To prove b), we first observe that $E(\sum f(t,s,Js)) = \int f(t,s,x)M(ds,dx)$ so that convergence of the integral is certainly sufficient for $\sum_s f(t,s,Js) < \infty$ a.s. On the other hand, suppose $\int f(t,s,x)M(ds,dx) = \infty$. We then compute

$$E[\exp(-\sum_s f(t,s,Js))] \leq E[\exp(-\sum_{k=1}^{\infty} 2^{-k}N_k)] = \prod_{k=1}^{\infty} E[\exp(-2^{-k}N_k)]$$
$$= \prod_{k=1}^{\infty} \exp\{M(B_k)[\exp(-2^{-k}) - 1]\} = \exp\{\sum_{k=1}^{\infty} [\exp(-2^{-k}) - 1]\, M(B_k)\}\ .$$

However, $e^{-x} - 1 \sim -x$ as $x \to 0$ so that $\sum_k [\exp(-2^{-k}) - 1]\, M(B_k) > -\infty$ if and only if $\sum_{k=1}^{\infty} 2^{-k}M(B_k) < \infty$. Since $\infty = \int f(t,s,x)M(ds,dx) \leq \sum_{k=1}^{\infty} 2^{-k+1}M(B_k)$ it follows that $E[\exp(-\sum_s f(t,s,Js))] = 0$, i.e., $\sum_{s\in[0,T]} f(t,s,Js) = \infty$ a.s.

3. Variational Sum Approximation

In this section we consider the approximation of $Y(t)$ by variational sums $V_t(\Pi,f,X) = \sum_j f(t,s_j,\Delta X_j)$, where $\Pi = \{s_0 < s_1 < \ldots < s_n\}$ is a partition of $[0,T]$ and $\Delta X_j = X(s_{j+1}) - X(s_j)$ is the observed increment of the process X over the partition interval (s_j,s_{j+1}). Consideration of the limiting behavior of $V_t(\Pi,f,X)$ as $||\Pi|| \to 0$ is natural in light of the likelihood ratio representation for additive process, and also the representation of linear processes as stochastic integrals. In general, we would like to know in what manner the Stieltjes sums converge to a stochastic integral. Throughout we assume that for each t, $f(t,\cdot,\cdot)$ is continuous on $[0,T]XR$.

The "strong variation" $W_t(f,X)$ is defined by $W_t(f,X) = \sup V_t(\Pi,f,X)$ where the supremum is over all partitions of $[0,T]$. We shall also consider limits of $V_t(\Pi,f,X)$ as $||\Pi|| = \max(s_{j+1} - s_j) \to 0$. In general $\lim_{||\Pi||\to 0} U(\Pi)$ denotes the limit along a net directed by partitions partially ordered by mesh. Thus $\lim_{||\Pi||\to 0} U(\Pi)$ exists if and only if $\lim\sup_{\delta\downarrow 0\ ||\Pi||<\delta} U(\Pi) = \lim\inf_{\delta\downarrow 0\ ||\Pi||<\delta} U(\Pi)$ and both are finite.

We shall refer to the following consequence of the sample path properties of additive process as the truncation principle. It has been used by many authors, but it is convenient to formulate it explicitly here.

Truncation Principle Let $f(t,s,x)$ be a Borel measurable function which is right continuous in s, and continuous a.e. $[M_T]$ as a function of x. If $\pm\varepsilon$ are not atoms of M_T, then

$$\lim_{||\Pi||\to 0} \sum_j f(t,s_j,\Delta X_j) I_{[|\Delta X_j|>\varepsilon]} = \sum_s f(t,s,Js) I_{[|Js|>\varepsilon]}$$

The following lemma will be useful for analysis of limiting properties of $V_t(\Pi,f,X)$. A version was noted in a remark of Fristedt and Pruitt (1972) p.65.

Lemma 1 (Extended Kolmogorov zero-one principle) Let $\{X_k(t),\ t \in I,\ k \geq 1\}$ be a sequence of independent stochastic processes. Let $F_n = \sigma\{Y_k(t),\ t \in I,\ k \geq n\}$

and $\mathcal{F} = \bigcap_{n=1}^{\infty} \mathcal{F}_n$. Then $A \in \mathcal{F}$ implies $P(A)$ is zero or one.

Proof Let $\varepsilon > 0$ be arbitrary, and suppose $A \in \mathcal{F}$. Since $\mathcal{F}_n$ is generated by the algebra consisting of finite disjoint unions of sets of the form $\bigcap_{j=1}^{m} [Y_{k_j}(t_j) \in A_j]$, $k_j \geq n$, it follows that there exists a set of the above form with $P(A \Delta B) < \varepsilon$. The remainder of the proof follows by standard arguments (see, e.g., Breiman (1968) page 40).

We now apply this lemma to the strong variation process.

Theorem 2 $|W_t(f,X) < \infty$ a.s. or $|W_t(f,X)| = \infty$ a.s.

Proof Let $X^{(\varepsilon)}(t)$ be the process derived from $X(t)$ by deleting all jumps of magnitude greater than ε, i.e., $X^{(\varepsilon)}(t) = X(t) - \Sigma \{Js: |Js| > \varepsilon\}$, and let $\{\delta_k\}$ be a sequence with $\delta_k \to 0$, and $\pm\delta_k$ not atoms of M_T. From the sample path properties of X it follows that $|W_t(f,X)| < \infty$ iff $|W_t(f,X^{(\delta_k)})| < \infty$ for every k.

Let $X_k(t) = X^{(\delta_{k-1})}(t) - X^{(\delta_k)}(t) = \Sigma \{Js: \delta_k \leq |Js| < \delta_{k-1}\}$. The processes X_k are independent, and the random part of $X^{(\delta_n)}$ is a function of $\{X_k(t),\ t \in [0,T],\ k \geq n\}$ and hence $X^{(\delta_n)}$ is measurable with respect to $\mathcal{F}_n = \sigma\{X_k(t),\ t \in [0,T],\ k \geq n\}$. Thus, for each n $[|W_t(f,X)| < \infty] = [|W_t(f,X^{(\delta_n)})| < \infty] \in \mathcal{F}_n$, so the above event is a tail event, which by lemma 1 must have probability zero or one.

We may also apply the zero-one principle to obtain a characterization of convergence of variational sums. The following result extends the convergence-divergence dichotomy observed by several authors for functions $f(t,s,x) = |x|^p$ to the more general framework.

Theorem 3 For any $t \in [0,T]$

$P[\lim_{||\Pi|| \to 0} V_t(\Pi,f,X)$ exists$]$ is zero or one. Moreover, if $\lim_{||\Pi|| \to 0} V_t(\Pi,f,X)$

exists, and $\int_{[0,T]\times[-1,1]} |f(t,s,x)|M(ds,dx) < \infty$ (so that the limit is finite).

then the limit is of the form $\lim_{||\Pi||\to 0} V_t(\Pi,f,X) = C + Y(t)$ where $Y(t) = \Sigma\, f(t,s,Js)$,

and C is a constant.

Proof Choose $\pm\delta$ not to be an atom of M_T. A truncation argument shows that

$$\lim_{||\Pi||\to 0} V_t(\Pi,f,X) = \Sigma\,\{f(t,s,Js): |Js| > \delta\} + \lim_{||\Pi||\to 0} V_t(\Pi,f,X^{(\delta)})$$

provided either side exists. Hence, $\lim_{||\Pi||\to 0} V_t(\Pi,f,X)$ exists if $\lim_{||\Pi||\to 0} V_t(\Pi,f,X^{(\delta)})$

exists for all δ. The method of proof used in Theorem 2 shows that the event

$[\lim_{||\Pi||\to 0} V_t(\Pi,f,X)$ exists] is a tail event and so has probability zero or one.

Now assume $\lim_{||\Pi||\to 0} V_t(\Pi,f,X)$ exists a.s. and suppose $\int_{[0,T]\times[0,1]} |f(t,s,x)|M(ds,dx) < \infty$

so that according to Theorem 1, $\Sigma\, f(t,s,Js)$

converges absolutely with probability one. Let $D = \lim_{||\Pi||\to 0} V_t(\Pi,f,X) - \Sigma f(t,s,Js)$.

By the above it follows that D is finite and

$D = \lim_{||\Pi||\to 0} V_t(\Pi,f,X^{(\delta)}) - \Sigma\,\{f(t,s,Js): |Js| \leq \delta\}$. Again the method of proof

of Theorem 2 shows D is tail measurable, and hence must be constant a.s. This

completes the proof.

It should be noted that the constant C need not be zero (i.e., the variational sum need not converge to what we would like it to). An example of Fristedt and Taylor shows that for $f(t,s,x) = x^2$ different limits can occur (see example 1, page 275, of Fristedt and Taylor (1973)). In the next section we consider a further specialization of f sufficient to ensure convergence to the desired limit. It will include the linear processes, and also many important likelihood ratios.

§4. Applications to Linear Processes and Signal Detection

In this section we shall consider functions of the form $f(t,s,x) = g(t,s)h(x)$ where $g(t,s)$ is uniformly bounded, and $h(x) = O(|x|^\gamma)$ for some $\gamma > 0$. By Theorem 1, the desired variational sum limit $\Sigma\, f(t,s,Js)$ will converge if and

only if $\int_{[0,T]\times[-1,1]} f(t,s,x)M(ds,dx) < \infty$, so we shall assume $\int_{-1}^{1} |x|^{\gamma} M_T(dx) < \infty$, and $\gamma < 2$.

In order to prove the theorems of this section we will need inequalities due to Millar and to Bretagnolle. It is convenient to state these results here. The Millar inequalities, Theorem 2.1 of [23], are not stated exactly as in his paper but are easily obtained from it by a careful reading of his proof and taking $X = Y(1)$ where $Y(t)$ has stationary independent increments.

Theorem (Millar): Let X be an infinitely divisible random variable with characteristic function $\exp[itb + \int(e^{itx} - 1 - \frac{itx}{1+x^2})M(dx)]$. Suppose that the support of Lévy measure, M is contained in $[-a,a]$, $0 \le a \le \infty$, and $\int_{|x|\le a} |x|^{\gamma} M(dx) < \infty$

(i) If $\alpha \ge 1$ then $E|X|^{\alpha} \le 2\int_{|x| \le a} |x|^{\alpha} M(dx) + 2^{\alpha-1}|EX|^{\alpha}$

(ii) If $\alpha \le 1$ and if $b = \int_{|x| \le a} \frac{x}{1+x^2} M(dx)$, then $E|X|^{\alpha} \le \int_{|x| \le a} |x|^{\alpha} M(dx)$.

We will combine Millar's inequalities with those of Bretagnolle's Theorem 2, [4], stated here in the restricted form which we will actually use.

Theorem (Bretagnolle) Let $X_1,\ldots,X_n$ be independent random variables with $EX_i = 0$, $1 \le i \le n$. Let $S(k) = \Sigma_1^k X_i$ and let $\mathcal{P}$ be the class of all finite subsets $\Pi = \{n_0,n_1,\ldots,n_k\}$, $0 = n_0 < n_1 < \ldots < n_k = n$ of $\{0,1,\ldots,n\}$. Then for $1 < p < 2$,

$$E\Sigma_1^n |X_i|^p \le \sup_{\mathcal{P}} \Sigma_{n_i \in \Pi} E|S(n_i) - S(n_{i-1})|^p \le C_p \Sigma_1^n E|X_i|^p$$

where C_p is a constant depending only on p.

The case $p \le 1$ of Theorems 4 and 5 below are proved in Hudson and Mason (1976). We will give the proofs for $p > 1$. C_p denotes a constant depending only on p. Theorem 4 below extends the results of Theorem 3 of Bretagnolle [4]. We shall write $W_g(p,X)$ and $V_g(\Pi,p,X)$ for $W_t(f,X)$ and $V_t(\Pi,f,X)$ when $f(t,s,x) = g(t,s)|x|^p$. We shall assume from here on that $a(t)$ is of bounded variation if $p \ge 1$, and $\int X/(1+X^2)\, M_t(dX) < \infty$ if $p < 1$.

Lemma 2. Suppose $X(t) \sim (a(t),M_t)$ and $0 < p < 2$ and $M_T([-r\ \ ]^c) = 0$ for some

$a > 0$. If $E(X(t)) = 0$, then

$$E(W_g(p,X)) \leq C_{pg} \int |x|^p M_T(dx).$$

Proof. Let P_n denote the class of all partitions of $[0,T]$ consisting of n^{th} order binaries. Let $Y_n = \{\sup V(\Pi,|x|^p,X): \Pi \in P_n\}$. By the sample path properties of X it follows that $W_g(p,X) \leq C_g W_1(p,X)$ and $Y_n \uparrow W_1(p,X)$ a.s. Let n be fixed and let $X_j = X(j2^{-n}) - X((j-1)s^{-n})$, $1 \leq j \leq 2^n T$. By Bretagnolle's inequality we have $E(Y_n) \leq C_p \sum_j E|X_j|^p$ for some constant C_p depending only on p. Since $E(X_j) = 0$, we may use the inequalities of Millar to obtain $\sum_j E|X_j|^p \leq \sum_j \{2 \int |x|^p (M_{t_j}(dx) - M_{t_{j-1}}(dx))\} = 2 \int |x|^p M_T(dx)$. Thus $E(Y_n) \leq 2C_p \int |x|^p M_T(dx) < \infty$. The monotone convergence theorem shows $W_1(p,X) < \infty$ and then this completes the proof.

Theorem 4. If $X \sim (a(t),M_t)$ is a general process with independent increments, $W_g(p,X) < \infty$ a.s. if and only if $\int_{|x|<1} |x|^p dM_T(x) < \infty$.

Proof. To see that $\int_{|x|<1} |x|^p M_T(dx) < \infty$ implies $W_g(p,X) < \infty$ a.s. for X as above, we use truncation, Lemma 2 and the fact that $W_g(p,X) < \infty$ a.s. if and only if $W_1(p,X) < \infty$. Write $Y_1(t) = X^{(1)}(t) - f^{(1)}(t)$ where $f^{(1)}(t)$ is defined by $f^{(1)}(t) = a(t) - \int_{|x|>1} \frac{x}{1+x^2} M_t(dx) - \int_{|x|<1} (\frac{x}{1+x^2} - x) M_t(dx)$. Since X(t) is stochastically continuous, $f^{(1)}(t)$ is continuous. Also $f^{(1)}(t)$ is of bounded variation since for any Borel set A, $M_t(A)$ is a nondecreasing function of t. The characteristic function of $Y_1(t)$ is $\exp[\int_{|x|<1} (e^{iux} - 1 - iux) M_t(dx)]$ and so $EY_1(t) = 0$. By Lemma 2, $W_g(p,Y_1) < \infty$ a.s. Hence $W_g(p,X^{(1)}) < \infty$ a.s. Let N denote the number of jumps of X(t), $0 \leq t \leq T$, of absolute value greater than one. The sample path properties of the X(t) process show that with

probability one the paths are bounded, by, say $B = \sup\{|X(t)|: 0 \le t \le T\} < \infty$. Then for any partition Π, there are at most N terms of $V(\Pi,|x|^p,X)$ which do not occur in $V(\Pi,|x|^p,X^{(1)})$. (The increments of X(t) and of $X^{(1)}(t)$ are the same unless an interval contains a jump of absolute value greater than one.) Now N is a Poisson random variable with expectation $\int_{|x|\ge 1} dM_T$. Thus for any partition Π of [0,T], $V(\Pi,|x|^p,X) \le V(\Pi,|x|^p,X^{(1)}) + 2NB^p$. That is,

$$W_1(p,X) \le W_1(p,X^{(1)}) + 2NB^p < \infty \text{ a.s.}$$

It remains to show that if $\int_{|x|<1} |x|^p M_T(dx) = \infty$, then $W_1(p,X) = \infty$ a.s. As a corollary to the Truncation Principle we observe that $\Sigma\{|J(s)|^p: 0 < s \le T\} \le \lim_{||\Pi||\to 0} \inf V(\Pi,|x|^p,X) \le W_1(p,X)$. But by Lemma 2, $\int_{|x|<1} |x|^p M_T(dx) = \infty$ implies that $\Sigma\{|J(s)|^p: 0 < s < T\} = \infty$ a.s. By the above inequality, $W_1(p,X) = \infty$ a.s., which implies $W_g(p,X) = \infty$ a.s.

The following theorem is well known for nested sequences of partitions. See Kallenberg (1974), or Hudson and Mason (1976).

Theorem 5. If $X(t) \sim (a(t),M_t)$, $0 < p < 2$, and $f(t,s,x) = g(t,s)|x|^p$ with g bounded, then $\lim_{||\Pi||\to 0} V_t(\Pi,f,X) = \Sigma f(t,s,Js)$ a.s. The limit is finite if $\int_{|x|<1} |x|^p M_T(dx) < \infty$.

Proof. Let $g^{(\varepsilon)}(t) = a(t) - \int_{|x|>\varepsilon} x/(1+x^2)M_t(dx) - \int_{|x|\le\varepsilon} (\frac{x}{1+x^2} - x)M_t(dx)$. By the truncation principle, and path properties we have for $\pm\varepsilon$ not atoms of M_T

$$\begin{aligned}
\overline{\lim_{||\Pi||\to 0}} V_t(\Pi,f,X) &\le \overline{\lim_{||\Pi||\to 0}} \Sigma\{f(t,s_j,\Delta X_j): |\Delta X_j| < \varepsilon\} \\
&\quad + \overline{\lim_{||\Pi||\to 0}} \Sigma\{f(t,s_j,\Delta X_j): |\Delta X_j| \ge \varepsilon\} \\
&\le \Sigma\{f(t,s,Js): |Js| \ge \varepsilon\} \\
&\quad + C_g \overline{\lim_{||\Pi||\to 0}} V(\Pi,|X|^p,X^{(\varepsilon)})
\end{aligned}$$

where $X^{(\varepsilon)}(t) = X(t) - \Sigma\{J(s): |J(s)| \geq \varepsilon\}$. Let $\varepsilon \downarrow 0$ to get

$$(4.1) \qquad \overline{\lim_{||\Pi||\to 0}} V(\Pi, f, X) \leq \Sigma f(t,s,Js) + C_g \overline{\lim_{\varepsilon\downarrow 0}} \overline{\lim_{||\Pi||\to 0}} V(\Pi, |X|^p, X^{(\varepsilon)}).$$

We will show that $\overline{\lim_{\varepsilon\downarrow 0}} \overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, X^{(\varepsilon)}) = 0$ a.s. Now by the C_r inequality,

$$(4.2) \quad V(\Pi, |x|^p, X^{(\varepsilon)}) \leq 2^{p-1} V(\Pi, |x|^p, f^{(\varepsilon)}) + 2^{p-1} V(\Pi, |x|^p, Y^{(\varepsilon)}) \text{ where } Y^{(\varepsilon)}(t)$$

$= X^{(\varepsilon)}(t) - g^{(\varepsilon)}(t)$. Note that $Y^{(\varepsilon)}(t)$ satisfies the conditions of Lemma 2 so that $EW_1(p, Y^{(\varepsilon)}) \leq C_p \int_{|x|<\varepsilon} |x|^p M_T(dx)$. Now for $p > 1$, the fact that $g(t) = \lim q^{(\varepsilon)}_{(t)}$ is continuous and of bounded variation implies that for any $\varepsilon > 0$

$$(4.3) \qquad \overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, g^{(\varepsilon)}) = 0.$$

Thus

$$\overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, X^{(\varepsilon)}) \leq 2^{p-1} \overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, Y^{(\varepsilon)}).$$

But for $\delta > \varepsilon$

$$\begin{aligned}
\overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, Y^{(\varepsilon)}) &\leq 2^{p-1} \overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, Y^{(\delta)} - Y^{(\varepsilon)}) \\
&\quad + 2^{p-1} \overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, Y^{(\delta)}) \\
&\leq 2^{p-1} \Sigma\{|J(s)|^p: \varepsilon < |J(s)| \leq \delta\} \\
&\quad + 2^{p-1} W_1(p, Y^{(\delta)}) .
\end{aligned}$$

Thus, $E[\overline{\lim_{\varepsilon\downarrow 0}} \overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, Y^{(\varepsilon)})] \leq 2^{p-1} EW_1(p, Y^{(\delta)}) + 2^{p-1} \int_{\varepsilon<|x|\leq\delta} |x|^p M_T(dx)$.

$$\leq 2^{p-1} \int_{\varepsilon<|x|\leq\delta} |x|^p M_T(dx) + 2^{p-1} C_p \int_{|x|<\delta} |x|^p M_T(dx) \to 0 \text{ as } \delta\downarrow 0.$$

Since $V(\Pi, |x|^p, Y^{(\varepsilon)}) \geq 0$, this shows that

$$(4.4) \qquad \overline{\lim_{\varepsilon\downarrow 0}} \overline{\lim_{||\Pi||\to 0}} V(\Pi, |x|^p, Y^{(\varepsilon)}) = 0 \text{ a.s.}$$

By (4.2), (4.3), and (4.4), we get $\overline{\lim_{\varepsilon \downarrow 0}} \; \overline{\lim_{||\Pi|| \to 0}} \; V(\Pi, |x|^p, X^{(\varepsilon)}) = 0.$

But by (4.1), this implies that $\overline{\lim_{||\Pi|| \to 0}} \; V(\Pi, f, X) \leq \Sigma\{f(t,s,Js): 0 < s \leq T\}.$

As a corollary to the truncation principle we know that

$$\Sigma\{f(t,s,Js): 0 < s \leq T\} \leq \overline{\lim_{||\Pi|| \to 0}} \; V(\Pi, f, X)$$

and hence, $\lim_{||\Pi|| \to 0} V(\Pi, f, X) = \Sigma\{f(t,s,Js): 0 < s \leq T\}.$

Linear processes considered by Eastwood and Lugannani are particular cases of the results of the previous sections with $Y(t) = \int_a^b g(t,s)dX(s)$ $= \Sigma g(t,s)Js$. By Theorem 5 we observe that the natural variational (pathwise) approximations $\sum_{\Pi} g(t,s_j)\Delta X_j$ converge a.s. to Y(t). Our results show the convergence is almost sure for a class which includes linear processes. Consequently pathwise approximation can be done as outlined previously once the filter function g is known. Riedel (1980) also considers stochastic integrals with convergence being in probability. Our results also extend his.

As another application of the previous results, we can approximate the likelihood ratio of two processes with independent increments. An explicit formula for $\ln(dux_1/dux_2)$ for two additive processes has been given in Brockett, Hudson and Tucker (1978) as $\sum_{n=1}^{a} (\Sigma\{\ln \frac{dM}{dN}(s,Js): \; 0 \leq s \leq T, \; Js \in I_n\}$ $- \int_{[0,I]xI_n} (\frac{dM}{dN} - 1)dN\}$, where $I_n = \{x: \varepsilon_{n+1} < x \leq \varepsilon_n\}$, $\varepsilon_n \downarrow 0$ and M and N are the time-jump measures of the processes $X_1(t)$ and $X_2(t)$ respectively. For actual computation of dux_1/dux_2 given a sample function, some approximation must be made to this infinite sum. Using our results with $f(t,s,x) = \ln\frac{dM}{dN}(s,x)$,

we observe that an a.s. consistent estimate can be obtained by using $V(\Pi,f,X_2)$. This gives an alternative approximation technique to that utilized by Stuck (1976) for assessing the likelihood ratio. Essentially he assumes no jumps of magnitude less than ε occur, and then calculates the finite summation $\Sigma\{\ln\frac{dM}{dN}(s,Js)\}$ exactly for these approximating processes. Here we use the entire sample function for the untruncated process, but use a variational sum approximation for the contribution of the small jumps.

References

1. Berman, S. M., Sign-invariant random variables and stochastic processes with sign invariant increments, Trans. Amer. Math. Soc. 119, 216-243, MR 32 #313 (1965).

2. Blumenthal, R. M. and R. K. Getoor, Sample functions of stochastic processes with stationary independent increments, J. Math Mech. 10, 493-516, MR 23 #A689 (1961).

3. Breiman, L., Probability, Addison-Wesley, Reading, Massachusetts, 1968.

4. Bretagnolle, J., p variation de fonctions aléatoires. 2ième partie: Processus à accroissements indépendants, Séminaire de Probabilitiés VI, Université de Strasbourge, Lecture Notes Springer-Verlag (1972).

5. Brockett, Patrick L., "Variational sums of infintesimal systems," Z.Wahr. 38, 293-307 (1977).

6. Brockett, Patrick L., and Tucker, Howard G., "A conditional dichotomy theorem for stochastic processes with independent increments,"J. Multi Analysis 7, 13-27 (1977).

7. Brockett, Patrick L., W. N. Hudson, and Howard G. Tucker, "The distribution of the likelihood ratio for additive processes," J. Multi Analysis 8, 233-243 (1978).

8. Cogburn, R., and H. G. Tucker, A limit theorem for a function of the increments of a decomposable process, Trans. Amer. Math. Soc. 99, 278-284 (1961).

9. Eastwood, Lester F., Jr., and Robert Lugannani, "Approximate likelihood ratio detectors for linear processes," IEEE Trans. Inf. Th. IT-23, 482-489 (1977).

10. Fristedt, B., Sample function behavior of increasing processes with stationary independent increments, Pacific J. Math. 21, 21-33 (1967).

11. Fristedt, B., and W. E. Pruitt, Uniform lower functions for subordinators, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 24, 63-70 (1972).

12. Fristedt, B., and S. J. Taylor, Strong variation for the sample functions of a stable process, Duke Math. J. 40, 259-278 (1973).

13. Greenwood, P. E., The variation of a stable path is stable. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 14, 140-148 (1969).

14. Greenwood, P. E., and B. Fristedt, Variations of processes with stationary independent increments, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 23, 171-186 (1972).

15. Hudson, W. N., and J. D. Mason, Variational sums for additive processes, Proc. Amer. Math. Soc. 55, 395-399 (1976).

16. Hudson, W. N., and H. C. Tucker, Limit theorems for variational sums, Trans. Amer. Math. Soc. 191, 405-426 (1974).

17. Kallenberg, O., Path properties of processes with independence and interchangeable increments. Z. Wahrscheinlichkeitstheorie and Verw. Gebiete 28, 257-271 (1974).

18. Lugannani, R., and J. B. Thomas, "On a class of stochastic processes which are closed under linear transformations," Information and Control 10, 1-21 (1967).

19. D. Middleton, "A Statistical Theory of Reverberation and Similar First Order Scattered Fields. Part I: Waveform and the General Process", IEEE Trans. Inf. Theory 13, 372-392 (1967).

20. D. Middleton, "A Statistical Theory of Reverberation and Similar First Order Scattered Fields. Part II: Moments, Spectra, and Spacial Distributions", IEEE Trans. Inf. Theory 13, 393-414 (1967).

21. D. Middleton, "A Statistical Theory of Reverberation and Similar First Order Scattered Fields. Part III: Waveforms and Fields", IEEE Trans. Inf. Theory 18, 35-67 (1972).

22. D. Middleton, "A Statistical Theory of Reverberation and Similar First Order Scattered Fields. Part IV: Statistical Models", IEEE Trans. Inf. Theory 18, 68-90 (1972).

23. Millar, P. W., Path behavior of processes with stationary independent increments, Z. Wahrscheinlichkeitstheorie and Verw. Gebiete 17, 53-73 (1971).

24. Monroe, I., On the γ-variation of processes with stationary independent increments, Ann. Math. Stat. 43, 1213-1220 (1972).

25. Riedel, M., Representation of the characteristic function of a stochastic integral, J. Appl. Prob. 17, 448-455 (1980).

26. Stuck, Bart, "Distinguishing stable probability measures, Part I: Discrete time, Part II: Continuous time, "Bell System Technical Journal 55, 1125-1196 (1976).

ARL-TP-82-34
2 August 1982

IDENTIFIABILITY FOR DEPENDENT MULTIPLE DECREMENT/COMPETING RISK MODELS

by

Barry C. Arnold*
Patrick L. Brockett**

* Department of Statistics, University of California, Riverside, California
** Department of Finance and Applied Research Laboratories, The University of Texas at Austin, Austin, Texas

This paper has been submitted to Scandinavian Actuarial Journal for publication.

This reasearch was supported in part by the Office of Naval Research under Contract N00014-81-K-0145.

§1 Introduction

In the study of multiple decrement lifetimes one considers a group of K lives. The group is said to survive at least as long as all members are living. In practice, the component lives are usually combined in annuity or insurance, or some other common undertaking, and frequently are related by blood, marriage or some joint undertaking which simultaneously exposes the individuals to common hazards and mortality risks. It follows that in general the joint survival function of the group should exhibit dependence between the component lives. The method of analysis now commonly used for multiple decrement analysis however assumes independence of the joint lives and joint life insurance payable upon the first death in the group is calculated using this assumption (c.f. Jordan 1975, Chapter 9).

The same mathematical model, but for a different purpose, is encountered in the theory of competing risks in biometry. Two very good reviews of the literature on this topic are Birnbaum (1979) and Moeschberger and David (1978). In this situation there are postulated to be K different mutually exclusive and collectively exhaustive causes or risks which compete to cause the eventual failure of the mechanism under study. The mechanism is assumed to function until at least one of the causes or risks precipitates system failure. The model is fairly general; different components in a complex electrical network may "compete" for equipment failure, different diseases may "compete" for a life, or different pathways may compete for a chemical reaction. Again, dependence of the risks is often present (e.g. systematic diseases).

The common mathematical formulation of the multiple decrement (competing risk) model which we shall investigate postulates a lifetime X_i for the i^{th} person (i^{th} risk factor acting alone). The random variables

$(X_1, X_2, \ldots, X_k)$ are called the net lives corresponding to the k postulated possible causes of system failures. The net lives may be potentially observable random variables as in the multiple decrement model, and in the competing risk formulation for complex electrical equipment, or they may be a theoretical mathematical construct, as in the case of several diseases competing for lives, or a survival/censorship model. The actual life is defined to be the observed time of system failure, $Z = \min_{1 \le i \le k} X_i$.

Usually another variable J, the cause of failure is also observed, $J = i$ if $Z = X_i$. The pair (Z,J) is called the identified minimum. Thus, if $S_{\underset{\sim}{X}}(\underset{\sim}{x}) = P[X_i > x_1, \ldots, X_k > x_k]$ is the joint survival function, then $S_Z(t) = S_{\underset{\sim}{X}}(t, \ldots, t)$ is observable, as are the so-called "crude life probabilities" $Q_i(t) = P[Z > t, J = i]$ $i = 1, 2, \ldots K$.

On the other hand, the joint survival function $S_{\underset{\sim}{X}}(\underset{\sim}{x})$ is not directly estimable. Knowledge of $S_{\underset{\sim}{X}}$ would completely determine the relationships between the competing risks (or insured lives in the multiple decrement life table). The identifiability problem, then, is to reconstruct $S_{\underset{\sim}{X}}$ from known $S_{\underset{\sim}{Z}}(t)$ and $Q_i(t)$, $i = 1, 2, \ldots, K$.

In section two we shall review the literature and previous contributions. In section 3 we introduce a bivariate model with dependence and marginal distributions of the Makeham type. This may be useful since the Makeham law adequately fits univariate life table date, and consequently this model provides a dependent alternative to the commonly used technique in multiple decrement life table analysis which assumes independent Makeham marginal distributions (c.f., Jordan 1975).

In section 4 we introduce dependent models which are scale mixtures of independent random variables, and prove identifiability. These models include dependent proportional hazards models, multivariate Pareto distributions, and also a multivariate logistic law.

§2. Review of the literature

The proofs of most of the results of this section can be found in the review papers by Birnbaum (1979) and Moeschberger and David (1978). These results are stated for completeness, and to put our contribution in context.

The first contribution to the subject was by D. Bernoulli (1760) who estimated mathematically the effect on the average life expectancy which would obtain if the risk factor small pox was eliminated from the population. Naturally, Bernoulli did not quite frame the problem as we have. In the modern guise, the first contribution to the identifiability problem was that of Berman (1963). He proved that if $X_1, X_2, \ldots, X_k$ are independent, then $F_{\underset{\sim}{X}}$ is uniquely identifiable from the knowledge of the identified minimum. In particular, he shows that

$$S_{\underset{\sim}{X}}(\underset{\sim}{x}) = \prod_{i=1}^{k} \exp\{-\int_0^{x_i} r_i(t)dt\}$$

where $r_i(t) = Q_i'(t)/S_Z(t)$ $i = 1,2,\ldots,k$. Thus, the usual actuarial practice of assuming independent lives for multiple decrement analysis can be based on identifiability.

For the case of dependent net lives much less has been done. Tsiatis (1975) showed that in the general case, $S_{\underset{\sim}{X}}$ is not uniquely identifiable from knowledge of the identified minimum. Thus, there can be several joint survival functions with the same values for $S_Z(t)$ and $Q_i(t)$, $i = 1,\ldots,k$. In particular he proves the following lemma.

Lemma 1 (Tsiatis 1975) For the competing risk model, $Q_i'(t) = \frac{dQ_i(t)}{dt} = \frac{\partial}{\partial x_i} S_{\underset{\sim}{X}}(\underset{\sim}{x}) \Big|_{x_1=x_2\ldots=x_k=t}$ at any point t at which the derivative exists.

It is clear from this lemma that the knowledge of the identified minimum can only determine $S_{\underset{\sim}{X}}$ and its derivatives along the line $X_1=X_2=\ldots=X_k=t$, and generally this is not sufficient to uniquely determine $S_{\underset{\sim}{X}}(\underset{\sim}{x})$. (Note that this is a variant of the famous Cauchy problem in partial differential equations, a topic of considerable study c.f. Copson [1975].)

Since the general dependent model cannot be identified uniquely, and since the independence model is often unrealistic (e.g. when studying systemic diseases, or when investigating the relationship between net lives), it is natural to assume $S_{\underset{\sim}{X}}$ belongs to a flexible parametric family and to attempt to establish identifiability within this family. This parametric approach has the advantage of requiring no information other than the observable distributions $S_Z(t)$ and $Q_i(t)$ $i = 1,\ldots,k$, and, of course, the parametric family involved. Unfortunately, not every popular multivariate family of distributions is identifiable (see e.g. example 2.2).

The next step in the development of this subject then is the formulation of identifiable dependent multivariate models. The development of this aspect of the theory has progressed in a somewhat ad-hoc manner, as the following examples show. The first three models are bivariate exponential models.Which bivariate extension is desired depends upon which of the properties of univariate exponentials we wish to preserve. Almost all the models assume there can be no ties for the cause of death variable J, i.e., failure is due to a simple cause.

Example 2.1 Assume

$$S_{\underset{\sim}{X}}(x_1,x_2) = \exp\{-c_1x_1-c_2x_2-c_3x_1x_2\}, \quad x_1,x_2 > 0$$

where c_1, c_2, c_3 are non-negative. Then $S_{\underset{\sim}{X}}(\underset{\sim}{x})$ is uniquely determined by S_Z, Q_1 and Q_2. This is the simplist bivariate exponential

Example 2.2 (Block-Basu Bivariate Exponential)

Assume $S_{\underset{\sim}{X}}$ is given by its density

$$f(x_1,x_2) = \begin{cases} \frac{Cc_1(c_2+c_3)}{c_1+c_2} \exp\left\{-c_1x_1-(c_2+c_3)x_2\right\} & x_1 < x_2 \\ \frac{Cc_2(c_1+c_3)}{c_1+c_2} \exp\left\{-(c_1+c_3)x_1-c_2x_2\right\} & x_1 > x_2 \\ 0 \text{ for } x_1 \text{ or } x_2 < 0 \end{cases}$$

where $C = c_1+c_2+c_3$, $c_1,c_2 > 0$, $c_3 \geq 0$

Then $S_{\underset{\sim}{X}}$ is not identifiable (see Basu and Ghosh (1978)). This model retains the exponential "lack of memory" property, but does not have exponented marginals.

Example 2.3 (Downton Bivariate Exponential)

Assume $S_{\underset{\sim}{X}}$ is given by its density function

$$f(x_1,x_2) = \frac{c_1c_2}{1-} I_0\left(\frac{2\sqrt{\rho c_1c_2x_1x_2}}{1-\rho}\right) \exp\left\{-\frac{c_1x_1+c_2x_2}{1-\rho}\right\}$$

for $x_2x_2 \geq 0$, $c_1,c_2 > 0$, $0 \leq \rho \leq 1$ where I_0 is the modified Bessel function of the first kind of order 0. Then $S_{\underset{\sim}{X}}$ is identifiable. (See Downton 1970). It has exponential marginals, and arises from a non-fatal shock model.

Example 2.4 (Bivariate Normal)

Assume that $S_{\underset{\sim}{X}}$ is given by the density

$$f(x_1x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}}$$

then $S_{\underset{\sim}{X}}$ is identifiable. See Nadas (1971) for a proof, and Basu and Ghosh (1978) for further comments.

We may also account for simultaneous causes of failure by the following mathematical mechanism. For each vector $s \in \{0,1\}^k$ let $Q_{\underline{s}}(t) = P(Z>t, \underline{J} = \underline{s})$, i.e. $Q_{\underline{s}}$ is the probability that the life exceeded t and the cause was simultaneously $i_1, \ldots, i_m$ where the i_j's are the coordinates of $\underline{s}$ equal to 1. In this framework we can identify 2^k-1 "causes" of death. In this setting the Marshall-Olkin k dimensional distribution given in the next example will corresponding to 2^k-1 independent exponential competing risks. This is then identifiable from Berman's results.

Example 2.5 (Marshall-Olkin bivariate exponential)

Assume

$$S_{\underline{X}}(x_1,x_2) = \exp -c_1x_1-c_2x_2-c_3 \max(x_1,x_2)$$

for $x_1,x_2 > 0$, where $c_1, c_2 > 0$ and $c_3 \geq 0$. Then $S_{\underline{X}}$ is identifiable.

The Marshall-Olkin exponential has exponential marginal distributions, and also the "lack of memory" property $S(x_1+t,x_2+t) = S(x_1,x_2)S(t,t)$. They also show that any bivariate distribution with both these properties cannot be alsolutely continuous.

These are the parametric classes which have been investigated so far, and the proofs of identifiability do not always yield constructive methods for parameter estimation. For example, the Downton exponential and the bivariate normal are shown to be identifiable by showing two distinct such distributions cannot have the same values for $S_Z(t)$ and $Q_i(t)$ i = 1, 2. Thus, although the parameters are known to be uniquely determined, no recipe is given for calculating these parameters. In the next two sections we introduce some parametric classes, prove identifiability and explicitly give parameter estimates.

§.3 A Dependent Bivariate Makeham Distribution

In his original paper of 1825, in which he derived the survival model which bears his name, Gompertz assumed "the average exhaustion of a man's power to avoid death to be such that at the end of equal infinitely small intervals of time he lost equal portions of his remaining power to oppose destruction which he had at the commencement of these intervals." Using the reciprocal of the hazard function as a measure of "ability to oppose destruction" gave a differential equation for the hazard function h(t) whose solution yields the survival function exp {-dexp(st)}. Gompertz recognized that death might be due to two causes, chance and deterioration, but Makeham (1860) was the one to make the requisite constant addition to the hazard function.

If we introduce dependence into a bivariate Makeham model by adopting a Marshall-Olkin approach we are led to the survival function

$$S(x,y) = \exp\{-c_1x-c_2y-c_3 \max(x,y) -d_1 \exp(s_1x) - d_2 \exp(s_2y)\} \qquad (3.1)$$

The rationale of this model is as follows. Let W_i denote the time to death of person i due to deterioration alone, a Gompertz random variable with parameters d_i and s_i, i = 1, 2. Random shocks or accidents occur according to Poisson processes. The intensity parameter of the process $N_i(t)$ of shocks which are fatal only to person i is c_i, i = 1, 2, and, due to the simultaneous exposure of the two persons, there is an intensity c_3 for random accidents like automobile crashs which could be fatal to both individuals simultaneously. Let $N_3(t)$ be this third Poisson process. All the random variables W_1, W_2, $N_1(t)$, $N_2(t)$ and $N_3(t)$ are assumed to be independent. Then if X is the actual life of person 1, and Y is the actual life of person 2, it follows that

$S(x,y) = P[X>x,Y>y] = P[W_1>x, N_1(x) = 0, W_2>y, N_2(y)=0, N_3(\max(x,y) = 0]$

$= \exp\{-d_1\exp(s_1x)\}\exp\{-c_1x\}\exp\{-d_2\exp(s_2y)\}\exp\{-d_2y\}\exp\{-c_3\max(x,y)\}$

which is (3.1).

Note that S(x,y) is indeed a bivariate Makeham distribution in the sense that the univariate marginal distributions are both Makeham distributions. In addition, if $d_1 = d_2$ and $s_1 = s_2$ (so the Gompertz "deterioration" components are the same for the two individuals), which is a common assumption in actuarial calculations, then even the observed minimum of X and Y has a Makeham distribution as well.

The next theorem shows that this bivariate Makeham is identifiable. Note that this does not follow from Berman's result since, for example if person 1 dies, we do not know whether deterioration or chance was at fault, and hence do not have an identified minimum for the formulation in terms of the independent variables W_1, W_2, N_1, N_2, N_3 even though it is an identified minimum problem in the X,Y formulation.

Theorem 1 The bivariate survival function (3.1) is identifiable.

Proof First we note that for this survival function which is not differentiable along the x = y line, we must modify the Tsiatis result (Lemma 1) to read

$$Q_1'(t) = \lim_{x\uparrow t} \left\{ \frac{\partial}{\partial x} S(x,t) \right\} \quad \text{and}$$

$$Q_2'(t) = \lim_{y\uparrow t} \left\{ \frac{\partial}{\partial x} S(t,y) \right\} .$$

The proof of this result is essentially the same as that of lemma 1. We calculate for (3.1).

$$Q_i'(t) = \{-c_i - d_i s_i \exp(s_i t)\} \quad S(t,t) \qquad i = 1, 2.$$

We note that

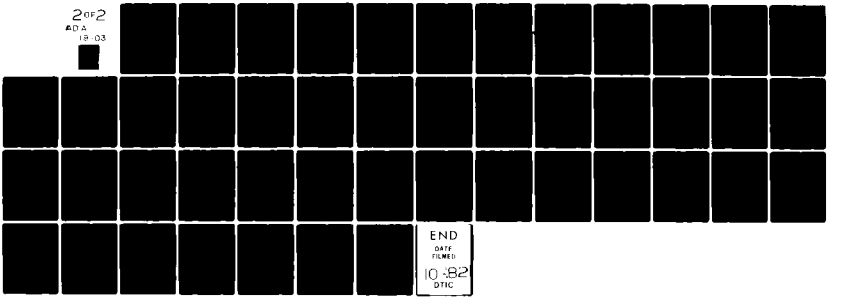
$$\psi_i(t) = \frac{d}{dt}\left[\frac{Q_i(t)}{S(t,t)}\right] = -d_i s_i^2 \exp(s_i t) \quad \text{is observable,}$$

and $\lim_{t\to\infty}\left\{\dfrac{\log(-\psi_i(t))}{t}\right\} = s_i$, $i = 1, 2$, so that

s_1 and s_2 are obtainable. Using the known s_i gives $d_i = \psi_i(0)/s_i^2$. Once s_i, d_i are known, c_i follows from $Q_i(t)/S(t,t)$. Now that $c_1, c_2, s_1, s_2, d_1, d_2$ are all known, c_3 follows from $S(t,t) = S_2(t)$ and all parameters have been uniquely determined.

AD-A119 103 TEXAS UNIV AT AUSTIN APPLIED RESEARCH LABS F/G 12/1
PREPRINTS RESULTING FROM WORK UNDER SRO 106: NON-GAUSSIAN SIGNA--ETC(U)
AUG 82 P L BROKETT N00014-81-K-0145
UNCLASSIFIED ARL-TP-82-29 NL

2 OF 2
ADA
119103



END
DATE FILMED
10-82
DTIC

§4. Scale mixture models.

Many parametric families of k-dimensional distributions can be characterized as scale mixtures of distributions with independent marginals. Thus we begin with $(U_1, U_2, \ldots, U_k)$ independent random variables and take $W > 0$ independent of the U_i's. The random vector $\underset{\sim}{X} = (X_1, \ldots, X_k)$ where $X_i = U_i/W$ has a distribution describable as a scale mixture of distributions with independent marginals. Such mixtures might arise when there is differential susceptibility to dying from different causes in a population, for example, if causes of death are closely related to diseases that are genetically controlled in a heterogeneous population. Not every such scale mixture will be identifiable. However, if we assume that the U_i's are independent with proportional hazards then we can get identifiability.

The following result generalizes the result of Elandt-Johnson (1976) and also Elandt-Johnson and Johnson (1980) §9.9 since we do not require the X_i's to have proportional hazard functions, but only that the $X_i|W$ have proportional hazards.

Theorem 2 Suppose conditional on $W = \lambda$, X_i are independent with proportional hazards $\delta_i h(t)$, so that the joint survival function is

$$S_{\underset{\sim}{X}}(\underset{\sim}{x}) = \int \prod_{i=1}^{k} \exp\{-\delta_i \lambda H(x_i)\} dG(\lambda) \text{ with } H(x) = \int_0^x h(t)dt.$$

Where G is the distribution function of random variable W.

(a) if h is known then $S_{\underset{\sim}{X}}$ is identifiable in $\{\delta_i\}$ and G.

(b) if G is known, then $S_{\underset{\sim}{X}}$ is identifiable in $\{\delta_i\}$ and h

Proof

(a) We first note that by changing the scale of G appropriately we may assume without loss of generality that $\delta_1 = 1$, and

$$S_X(x) = \int \exp\left\{-\lambda \sum_{i=1}^{k} \delta_i H(x_i)\right\} dG(\lambda)$$
$$= L_G\left(\sum_{i=1}^{k} \delta_i H(x_i)\right) \qquad (4.1)$$

where L_G denotes The Laplace-Stieltjes transform of the measure dG. Thus, from the above, and Lemma 1 we have for i = 1, ... , k.

$$Q_i'(t_0) = \frac{\partial S_{\underset{\sim}{X}}}{\partial x_i} \Big| x_1 = .. = x_k = t_0 = \delta_i h(t_0) \int \lambda \exp\left\{-\lambda H(t_0) \sum_{j=1}^{k} \delta_j\right\} dG(\lambda). \qquad (4.2)$$

Hence, $Q_i'(t_0)/Q_1'(t_0) = \delta_i$, i = 1, 2, ... , k. Now, with the knowledge of the δ_i's, and of $S_Z(t) = S_{\underset{\sim}{X}}(t)$, we know the Laplace-Stieltjes transform of G on the curve $H(t) \sum_{j=1}^{k} \delta_j$. Differentiability implies there is a limit point on this curve, so that the analyticity of the Laplace transform implies this determines G, and hence by integration with known $\{\delta_i\}$, we have determined $S_{\underset{\sim}{X}}$.

(b) Since G is assumed known, from (4.1) we have $L_G^{-1}(S_Z(t)) = H(t) \sum_{i=1}^{k} \delta_i$. As before, the δ_i's are uniquely determined from $Q_i'(t)/Q_i'(t)$, so the above formula determines H, and hence h.

Example 4.1 Suppose conditional on $W = \lambda$, $X_1, \ldots, X_k$ are independent Weibull variables with hazard $h(x_i) = c\delta_i \lambda x_i^{c-1}$, c known. Then unconditionally

$$S_{\underset{\sim}{X}}(\underset{\sim}{x}) = \int \prod_{i=1}^{k} e^{-\delta_i \lambda x_i^c} dG(\lambda) = L_G\left(\sum_{i=1}^{k} \delta_i x_i^c\right)$$

is an identifiable dependent competing risk model. (again G is the distribution function of W).

Example 4.2 Suppose the joint survival function of $(X_1, \ldots, X_k)$ is multivariate Pareto, i.e., has the survival function

$$S_{\underset{\sim}{X}}(\underset{\sim}{x}) = \left(\frac{1}{1 + \sum_{i=1}^{k} \frac{x_i - \mu_i}{\sigma_i}} \right)^{\alpha} \quad \text{for } x_i > \mu_i$$

(Such distributions are called Pareto (II) in Arnold (1981)) then $S_{\underset{\sim}{X}}$ is identifiable.

Proof We note first that $Q_i(t) = 0$ if and only if $t \leq \mu_i$, so the Q_i's determine the μ_i's. Knowing $\{\mu_i\}$ allows us to change variables. $\left(y_i = x_i - \mu_i\right)$ to reduce, without loss of generality, to the $\mu_i = 0$ case.
Now, the Laplace transform of a $\Gamma(\alpha,1)$ variable is $(1+u)^{-\alpha}$, so that by (4.1) $S_{\underset{\sim}{X}}(\underset{\sim}{x})$ represents the unconditional joint survival function corresponding to $X_1, \ldots, X_k$ being conditionally independent exponential variables with hazard functions wx_i/σ_i conditioned on W=w where W has a $\Gamma(\alpha,1)$ distribution. From (4.2) we have

$$Q_i'(t) = -\frac{\alpha}{\sigma_i} \left(1 + t \sum_{j=1}^{k} 1/\sigma_j\right)^{-\alpha-1}$$

so that

$$Q_i'(t)/S_Z(t) = \frac{1}{\sigma_i} \left[-\alpha \Big/ \left(1 + t \sum_{j=1}^{k} 1/\sigma_j\right) \right], \quad i = 1,2,\ldots k.$$

Let $\xi_i = Q_i'(0)/S_Z(0) = -\alpha/\sigma_i \quad i=1, \ldots, k$

and $\eta = Q_1'(1)/S_Z(1) = -\alpha \Big/ \left(\sigma_1 + \sum_{j=1}^{k} \sigma_1/\sigma_j\right)$ and we have

k+1 equations in k+1 unknowns which are easily solved for $\alpha, \sigma_1, \ldots, \sigma_k$. Thus the multivariate Pareto (II) is identifiable and an explicit representation for the parameters is given.

If the X_i's are known functions of random variables which, given W, are independent with proportional hazards, we may again conclude identifiability. As an illustration, we may consider scale mixtures of independent Weibull laws with known but arbitrary shape parameters. It turns out however that for scale mixtures of Weibull distributions, we do not necessarily have to know the shape parameters to conclude identifiability. Thus:

Theorem 3 Suppose that conditional upon W=w, $X_1, \ldots, X_n$ are independent Weibull variables with survival distributions $S_{X_i}(x_i) = \exp\{-\delta_i w x_i^{\alpha_i}\}$ respectively. Then the unconditional joint survival function is

$$S_{\underset{\sim}{X}}(\underset{\sim}{x}) = \int_0^\infty \prod_{i=1}^{k} \exp\{-\delta_1 w x_i^{\alpha_i}\}\, dG(w).$$

A sufficient condition that $S_{\underset{\sim}{X}}$ be uniquely determined by the knowledge of the identified minimum is that $E(W^\gamma) < \infty$ for some known $\gamma > 0$.

Proof First note that if $E(W^\gamma) < \infty$ for a known value of γ, then by taking γ^{th} roots of $WX_1, \ldots, WX_k$ we obtain a scale mixture of (new) Weibulls with a (new) scale mixing variable $W^{1/\gamma}$ which has a finite first moment. Thus it suffices to prove the theorem in the case $\gamma = 1$

From now on we assume $EW < \infty$. Knowledge of the identified minimum yields the functions $S_{\underset{\sim}{X}}(t,t,\ldots,t)$ and $Q_i(t)$, $i = 1, 2, \ldots, k$.

Now, $S_{\underset{\sim}{X}}(t,t,\ldots,t) = \mathcal{L}_G(\sum_{1=1}^{k} \delta_1 t^{\alpha_i})$, where $\mathcal{L}_G$ denotes the Laplace-Stieltjes transform of G and from Lemma 1.

$$Q_i'(t) = \int_0^\infty -\delta_i\alpha_i w x_i^{\alpha_i-1} \exp\{-w\Sigma\delta_i x_i^{\alpha_i}\}dG(w)\Big|_{x_1=..x_k=t}$$

$$= -\delta_i\alpha_i t^{\alpha_i-1} \int_0^\infty w \exp\{-w \sum_{i=1}^{k} \delta_i t^{\alpha_i}\}dG(w)$$

$$= \delta_i\alpha_i t^{\alpha_i-1} L'_G(\Sigma\delta_i t^{\alpha_i})$$

Consider

$$h_i(t) = \ell n \{tQ_i'(t)/ S_{\underset{\sim}{X}}(t,...,t) = \ell n\{\delta_i\alpha_i\} + \alpha_i \ell nt + \ell nk(t)$$

where $k(t) = L'_G(\Sigma\delta_i t^{\alpha_i})/L_G(\Sigma\delta_i t^{\alpha_i})$. Then $\lim_{t\to 0} h_i(t)/\ell nt = \alpha_i$ since $\ell nk(t)/\ell nt \to 0$. To see this note that $\ell nt \to -\infty$ as $t \to 0$ and $L'_G(\Sigma\delta_i t^{\alpha_i})/L_G(\Sigma\delta_i t^{\alpha_i}) \to \mu_G$, the mean of G, assumed finite.

Thus $\{\alpha_i\}$ can be recovered. Without loss of generality we may assume $\delta_1=1$ (since otherwise G may be rescaled accordingly). Now $\frac{Q_i'(1)}{Q_j'(1)} = \frac{\delta_i\alpha_i}{\delta_j\alpha_j}$, so the δ's may then be uniquely obtained. Knowing $\{\alpha_i\}$, and $\{\delta_i\}$ implies we know the Laplace transform of G along the curve $\sum_{i=1}^{k} \delta_i t^{\alpha_i}$. Since the Laplace transform is an analytic function, this determines G. Knowing G, $\{\delta_i\}$ and $\{\alpha_i\}$ uniquely determines $S_{\underset{\sim}{X}}(\underset{\sim}{x})$ in the class of mixtures, and exhibits identifiability.

Example 4.3 Suppose $(X_1,...,X_k)$ follow a multivariate Pareto (IV) law, i.e. their joint survival distribution is

$$S_{\underset{\sim}{X}}(\underset{\sim}{x}) = \left[1 + \sum_{i=1}^{k}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^{1/\gamma_i}\right]^{-\beta} \qquad x_i > \mu_i,\ i = 1,...,k$$

,

Then $S_{\underset{\sim}{X}}$ is identifiable.

Proof Since $Q_i(t) = 0$ if and only if $t \leq \mu_i$, we know μ_i, $i = 1,\ldots,k$ uniquely, and without loss of generality we may take $\mu_i = 0$.

Again, as in Example 4.2, we recognize the Laplace transform of the gamma law and observe that the multivariate Pareto (IV) is a gamma scale mixture of conditionally independent Weibull laws, with $\delta_i = 1/\sigma_i$ and $\alpha_i = 1/\gamma_i$ to fit the notation of the previous theorem. By Theorem 3 it follows that $S_{\underset{\sim}{X}}$ is identifiable. In fact we may estimate the exponent of the gamma scale mixing distribution quite readily since $\beta = -\ell n S_Z(t)/\ell n\left(1 + \Sigma\left(\frac{t}{\sigma_i}\right)^{1/\gamma_i}\right)$

The multivariate Pareto laws are particularly beautiful and useful multivariate laws for reliability and survival time modeling. In addition to the scale mixture of Weibull characterization above (see also Takahasi 1965), it is true that for such Pareto laws, the k-dimensional marginals of an n-dimensional Pareto are again Pareto for any $1 \leq k \leq n$, and also the conditional distribution of any k dimensional marginal given the remaining (n-k) variables is again Pareto. These are the types of properties which are also enjoyed by multivariate normals, but Pareto laws do not have the mathematical intractability and possible negative values for survival times.

Naturally, any known function of an identifiable competing risk model is also identifiable, so taking the logarithm of the crude life variables in the Pareto (III) distribution (i.e. P (IV) with $\beta = 1$) yields identifiability of the multivariate logistic distribution. (In fact the Pareto (III) law is sometimes referred to as the log-logistic distribution.) This is important since the logistic distribution function so closely resembles the normal distribution function. The lognormal distribution is not always tractable mathematically, however the Pareto often is easy to work with.

References

1. Arnold, Barry C. (1982). Pareto Distributions Monograph to be published by International Co-operative Publishing House.

2. Basu, A.P. and Ghosh, J. K. (1978). Identifiability of the multi-normal and other distributions under competing risk model. J. Mult. Anal., 8, 413-429.

3. Berman, S. M. (1963). Note on extreme values, competing risks and semi-Markov processes. Ann. Math. Statist., 34, 1104-1106.

4. Birnbaum, Z. W. (1979). On the mathematics of competing risks. DHEW Publication No. (PHS) 79-1351. U.S. Department of Health Education, and Welfare, pp. 1-58.

5. Block, H. W. and Basu, A. P. (1974). "A Continuous Bivariate Exponential Extension," J. Amer. Statist. Assoc. 69, pp. 1031-1037.

6. Copson, E. T. (1975). Partial Differential Equations, Cambridge University Press, Cambridge.

7. David. H. A. and Moeschberger, M. L (1978). The Theory of Competing Risks, Griffin's Statistical Monographs and Courses No. 39, MacMillan Publishing Co., Inc., New York.

8. Downton, F. (1970). "Bivariate Exponential Distributions in Reliability Theory," J. Roy. Statist. Soc. B 32, pp. 408-417.

9. Elandt-Johnson, R. C. (1976), Conditional failure time distributions under competing risk theory with dependent failure times and proportional hazard rates. Scand. Actu. J., 37-51.

10. Elandt-Johnson, R. C. and N. L. Johnson (1980) Survival Models and Data Analysis, John Wiley & Sons, New York.

11. Esary, J. D. and Marshall, A. W. (1974). "Multivariate Distributions with Exponential Minimums," Ann. Statist. 2, pp. 84-98.

12. Jordan, C. W. Jr. (1975). Life Contingencies. Society of Actuaries, Chicago.

13. Marshall, A. W. and Olkin, I. (1967). "A Multivariate Exponential Distribution," J. Amer. Statist. Assoc. 62, pp. 3-44.

14. Nadas, A. (1971). The distribution of the identified minimum of a normal pair determines the distribution of the pair, Technometrics, 13, 201-202.

15. Takahasi, K. (1965). Note on the multivariate Burr's distribution, Annals of the Institute of Statistical Mathematics, Tokyo, 17, 257-260.

16. Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. Proc. Nat. Acad. Ser., Wash., 72, 20-22.

ARL-TP-82-35
2 August 1982

THE UNDERWRITING RISK AND RETURN PARADOX REVISITED

by

Patrick L. Brockett*
Robert C. Witt**

* Department of Finance and Applied Research Laboratories, The University of Texas at Austin, Austin, Texas
** The Joseph H. Blades Centennial Memorial Professor of Insurance and Risk Management, Graduate School of Business, The University of Texas at Austin, Austin, Texas

This paper has been submitted to The Journal of Risk and Insurance.

This research supported in part by The Office of Naval Research under Contract N00014-81-K-0145.

THE UNDERWRITING RISK AND RETURN PARADOX REVISITED

Patrick L. Brockett and Robert C. Witt

Recently, Witt [1,2] suggested that one would expect to find a negative or inverse relationship between mean state loss ratios and standard deviations of the associated annual loss ratios by state for automobile insurance if underwriting risk (as measured by the standard deviation of loss ratios in a state) were properly included in insurance rates, other things being equal. However, based on an empirical test of this hypothesis, he found a direct rather than an inverse relationship between these variables. Hedges [3] later argued that the observed positive relationship could have resulted from a mathematical relationship between the average loss ratios and standard deviations and/or from certain institutional factors which affect loss ratios and insurance rates in the various states.

Based on some empirical tests of economic and institutional factors, Witt [4] showed that Hedges' institutional hypothesis could not be verified. His regression results for the whole industry showed a positive relationship between the underwriting risk and return variables even after he included various economic and institutional variables in the regression model. The regressions for the direct-writing, national-agency, and regional-speciality companies also revealed a positive relationship between the means and standard deviations of loss ratios by state. Thus, the addition of some institutional variables to his regression model did not reverse the significant positive relationship between mean loss ratios and standard deviations by state. Accordingly, he argued that the perverse observed positive correlation could not be easily explained by differences

in institutional factors among states.[1]

In regard to Hedges' mathematical hypothesis, Witt [4] argued that if the annual loss ratios by state are normally distributed there should be no relationship between the averages and the standard deviations of these ratios because there is no mathematical relationship between the mean and variance of a normally distributed random variable. Witt indicated that the Central Limit Theorem would suggest that the annual loss ratio in a state should be normally distributed because it is nothing more than the weighted average loss ratio of all the companies writing a given coverage in the state. Hence, he argued the positive correlations he observed could not be reasonably explained by the mathematical relationship postulated by Hedges.[2]

In Dr. Venezian's thoughtful comment, he correctly notes that the Cent Limit Theorem per se is not strong enough to imply that the sample mean and standard deviation (or variance) will become asymptotically uncorrelated, or independent, even though this is basically a unique characterisitic property of the normal limit law. However, in Witt's empirical work, a double Central Limit Theorem is in effect, and this effect is sufficient to imply that the correlation between sample means and standard deviations should be close to zero. Basically, this is the thrust of the heuristic argument given by Witt [4]. A precise mathematical justification for this argument will be given below.

We will also show under certain conditions that even treating premiums as random variables does not qualitatively change the results. This demonstration will help to amplify the remarks of Witt and Venezian. In the final section, we will explore other possible explanations, testable hypotheses and potentially promising avenues for further research on the issue of the proper relationship

---

[1]Witt [9, pp. 657-661].

[2]Witt [9, pp. 653-654].

between underwiting risk and return.[3]

Correlation and the Central Limit Effect

Let X_{kit} denote the loss on policy k in state i and year t and P_{kit} denote the corresponding premium. Here $k = 1,2,\ldots,N_{it}$; $i = 1,\ldots,51$; and $t = 1,\ldots,T$; where N_{it} is the total number of policies written by all insurers in year t in state i. The loss ratio for all insurers in state i during year t is as follows:

$$Z_{it} = \frac{\sum_{k=1}^{N_{it}} X_{kit}}{\sum_{k=1}^{N_{it}} P_{kit}}$$

As usual, $\bar{Z}_i = \sum_{t=1}^{T} Z_{it}/T$ and $S_i^2 = \sum_{t=1}^{T} (Z_{it} - \bar{Z}_i)^2/T$ are the temporal sample moments for state i for the T years analyzed. The discussion centers around the correlation of $(\bar{Z}_i, S_i^2)$ $i = 1,2,\ldots,51$.[4] To simplify notation, for an arbitrary random variable W, we shall let $\mu(W)$ denote the mean and, for $j \geq 2$, $\mu_j(W) = E[W - \mu(W)]^j$ denote the j^{th} central moment of W. As noted by Venezian, we have,

$$E(\bar{Z}) = \mu(Z)$$

$$\mu_2(\bar{Z}) = \frac{\mu_2(Z)}{T}$$

$$E(S_i^2) = \frac{T-1}{T}\mu_2(Z) \doteq \mu_2(Z)$$

$$\mu_2(S^2) = \frac{\mu_4(Z) - \mu_2^2(Z)}{T}$$

---

[3]The loss ratio in a state was used as an inverse measure of underwriting return by Witt [1,2].

[4]It should be noted that Witt [1,2] used loss ratio data for a series of years for all states and the District of Columbia, or 51 jurisdictions.

and most importantly,

$$\mathrm{Cov}(\bar{Z},S^2) = \frac{\mu_3(Z)}{T}$$

All of the above formulas include terms up to order 1/T, and can be found, for example, in Kendall and Stuart [5, pp. 243-246]. The main point is that all the second order moments including the covariance term are of order 1/T, and hence, when taking the correlation coefficient, the 1/T drops out, leaving the following.

$$(1) \qquad \mathrm{Corr}(\bar{Z},S^2) = \frac{\mathrm{Cov}(\bar{Z},S^2)}{\sqrt{\mu_2(\bar{Z})\mu_2(S^2)}} = \frac{\mu_3(Z)}{\sqrt{\mu_2(Z)\{\mu_4(Z) - \mu_2^2(Z)\}}}$$

Now, the point here is that $\mu_3(Z)$ should be close to zero because of a second central-limit effect. As shown above, the 1/T terms cancel. However, the dependence on N_{it} does not. Thus, this result does not depend in any way on the non-randomness of premiums, P_{ikt}, as suggested by Venezian.

Let us now calculate the ratio in (1) for the central moments of Z in terms of the central moments of the original loss and premium variables X and P.

The vehicle we will employ in this analysis is the multivariate version of Taylor's expansion, as specified in (2).

$$(2) \qquad f(x,y) = f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) + R$$

where

$$R = \frac{1}{2}\{f_{xx}(x_0,y_0)(x-a)^2 + 2f_{xy}(x_0,y_0)(x-a)(y-b) + f_{yy}(x_0,y_0)(y-b)^2\}$$

is the remainder term, and (x_0,y_0) is a point on the line segment joining (x,y) to (a,b). By dividing both the numerator and denominator of Z_{it} by N_{it}, we obtain

the ratio of average losses to average premiums:

$$Z_{it} = \frac{\overline{X}_{it}}{\overline{P}_{it}},$$

and by choosing $f(x,y) = x/y$; $a = \mu(\overline{X}) = \mu(X)$; $b = \mu(\overline{P}) = \mu(P)$, we obtain (after calculating the appropriate partial derivatives) the result specified in (3).

$$(3) \qquad Z_{it} = \frac{\mu(X)}{\mu(P)} + \frac{1}{\mu(P)}(\overline{X} - \mu(X)) - \frac{\mu(X)}{\mu^2(P)}(\overline{P} - \mu(P)) + R$$

where R is the random remainder term shown below:

$$R = -\frac{1}{P_0^2}(\overline{X}_i - \mu(X))(\overline{P}_i - \mu(P)) + \frac{2X_0}{P_0^3}(\overline{P}_i - \mu(P))^2$$

and (X_0,P_0) is a point between $(\overline{X},\overline{P})$ and $(\mu(X),\mu(P))$. Assuming fourth moments exist, one can show via Tchebychev's Inequality that $\sqrt{N_{it}} \cdot R$ converges in probability to zero as N_{it} increases without bound, and consequently from (3) that $\sqrt{N_{it}} \cdot Z_{it}$ is asymptotically normal for each (i.t). For convenience, we now drop the subscript (i,t).

Ignoring terms of higher order in 1/N yields the obvious approximation to (3) below.

$$(4) \qquad Z \doteq \frac{\mu(X)}{\mu(P)} + \frac{1}{\mu(P)}(\overline{X} - \mu(X)) - \frac{\mu(X)}{\mu^2(P)}(\overline{P} - \mu(P))$$

The moments of Z may now be read off from (4) and plugged into (1) after noting that $Z - E(Z) = \overline{W}$ where

$$W_k = \frac{1}{\mu(P)}(X_{ikt} - \mu(X)) - \frac{\mu(X)}{\mu^2(P)}(P_{ikt} - \mu(P))$$

$k = 1,2,\ldots,N_{it}$; and W_k has mean zero. Thus, we obtain

$$\mu_2(Z) = \mu_2(\overline{W}) = \frac{\mu_2(W)}{N}$$

$$\mu_3(Z) = \mu_3(\overline{W}) = \frac{\mu_3(W)}{N^2}$$

and $$\mu_4(Z) = \mu_4(\overline{W}) = \frac{\mu_4(W)}{N^3} + \frac{3N(N-1)}{N^4}\mu_2^2(W).$$ [5]

We observe that $\mu_3(Z)$ is of order $1/N^2$ and $\sqrt{\mu_2(Z)\{\mu_4(Z) - \mu_2^2(Z)\}}$ is of order $1/N^{3/2}$ so that the correlation coefficient in (1) is of order $1/\sqrt{N}$. Since the number of policy owners and companies per state is quite large, we would indeed expect the correlation of $\overline{Z}_i$ and S_i^2 to be close to zero, as Witt has implicitly claimed. Notice that the above derivation did not depend on the assumption of constant premiums, so the source of the positive correlation found in Witt [1,2] cannot be explained simply by differing premiums across companies.

Other Results, Hypotheses, and Directions for Further Research

Witt [1,2] noticed a cyclical pattern in loss ratios over time, and Venezian commented upon the autoregressive nature of the loss ratios. We wish to expand slightly upon Venezian's comment here. The problem is that the premiums in year t are not in fact independent of the past loss experience, and probably could be specified in the following way:

$$P_{ikt} = \lambda\mu_t + (1-\lambda)[\beta_1\overline{X}_{t-1} + \beta_2\overline{X}_{t-2} + \ldots + \beta_q\overline{X}_{t-q}] + E_t$$

---

[5] It might be noted that the coefficient of the second term in this equation could be written as $6\binom{N}{2}/N^4$.

That is, the premium is set in practice by taking some sort of a convex combination of the subjective expected loss and other economic factors for year t, μ_t, and a regression of the previously experienced losses for the last q years. An expense loading, E_t, is added to the loss component to obtain the gross premium. Frequently, q will be small and $\lambda = 0$. If $\lambda \neq 0$, then μ_t might reflect an effect on premium size due to investment income considerations and other prior economic knowledge that insurers recognize along with empirical loss and expense projections.

Empirical Bayes estimates for premium levels fit into this framework as well. In such situations, the loss ratios Z_{it} for $t = 1,2,\ldots,T$ are dependent random variables, even if $\overline{X}_{it}$ are independent. Indeed, if $\overline{X}_{i,t-1}$ is quite large, then $\overline{P}_{it}$ would tend to be large, and hence Z_{it} would tend to be small, other things being equal.

To formalize the above mathematically, we can use the Taylor series expansion from (4) and regression formulation for premium determination to obtain the following autoregressive formulation:

$$(6) \qquad Z_{it} = a_t + \sum_{j=1}^{q} b_j(\overline{X}_{t-j} - \mu(X))$$

The coefficients are determined by collecting the regression coefficients. Formula (6) offers a testable model for analysis. The techniques of Box and Jenkins [6] would seem applicable here for estimating parameters and determining the order. One might even use (6) for predicting future loss ratios and improving management planning and cash management.

Another approach to modeling the cyclical pattern is to utilize the results of Kemperman [7] on oscillating random walks. In this case, the loss ratios behave according to different random walks on either side of some specified line or real axis. While in the upper half of the plane, the random walk would have

a tendency to move down, and when in the lower half of the plane the loss ratio variable would have a propensity to move up. Thus, a cyclical pattern is an intrinsic part of the model. Related models of interest would include oscillating diffusion models studied by Keilson and Wellner [8] and the paper on records and oscillating random walks by Aggarwal [9].

Although Witt [1,2] used the temporal standard deviation associated with loss ratios in a state as a measure of underwriting risk, other possible measures of underwriting risk could be developed. For example, deviations from the auto-regressive patterns specified above could be used as a measure of underwriting risk, as Venezian has suggested. Recently, Witt and Miller [10] proposed another measure of underwriting risk based on the systematic component of total underwriting risk which basically captures the unpredictable and non-cyclical component in a state's loss ratio. They argued that the variability in the annual loss ratios in a state is a measure of total underwriting risk, which has two components: the systematic and unsystematic parts. The systematic component of underwriting risk was determined by regressing a state's loss ratio on the national loss ratio during a given time period. The resulting measure, referred to as "Beta," shows how the state's loss ratio varies with respect to the national loss ratio on a relative basis. The remaining variability, which was not explained by the national loss ratio, was designated as the unsystematic component of underwriting risk. In this guise, the model is statistically similar to the capital asset pricing model.

Witt and Miller [10] noted that systematic underwriting risk for a line of insurance cannot be eliminated by diversification across states. The unsystematic component of underwriting risk, in contrast, could be reduced or eliminated by writing insurance in more than one state. Since total underwriting risk would tend to decline as the number of states in which an insurer sold insurance

increased, they reasoned that increasing diversification by state would gradually eliminate the unsystematic risk faced by the insurer, leaving only systematic or national market risk. The remaining systematic risk, they argued, was associated with business risk or the unpredictable overall performance of the economy. Thus, they indicated that the systematic risk associated with a line of insurance could not be eliminated even if an insurer wrote in all states.[6]

Although the results were not reported in their paper, Witt and Miller [10] found a statistically significant positive relationship between average state loss ratios and their associated measure of systematic risk for 1971 through 1979. They expected that average state loss ratios would vary inversely with their "Betas," but they had to reject this hypothesis about underwriting risk and return for all company groups analyzed (direct writers, national-agency companies, regional-specialty companies, and all companies combined).[7] Thus, the use of a systematic measure of underwriting risk did not help to resolve the underwriting risk and return paradox. Perhaps, the development of a testable hypothesis including investment income along with underwriting income would help to resolve the paradox. It is possible that insurers were willing to accept relatively low underwriting returns during the 1971-1979 time period in exchange for higher anticipated investment returns. By keeping their prices

---

[6]They use the systematic measure of underwriting risk for determining the impact of rate regulation on loss ratios in states with different types of rate regulatory systems. Basically, they expected the systematic component of the variability in loss ratios for a state to be greater in highly regulated rate states than in states with competitive rate regulatory laws. They found some statistical support for this hypothesis. In general, the betas in competitive rate states were found to be lower than those in non-competitive ones. This result seemed to suggest that insurers faced greater systematic underwriting risk in highly regulated or non-competitive rate states than in competitive ones. It might be useful to perform a similar analysis using the cyclical auto-regressive formulation alluded to above.

[7]Statistical summaries of these results can be obtained from Witt.

relatively low, insurers would have incurred relatively higher loss ratios in order to improve their cash flow for investment purposes. A test of such a hypothesis over a longer time period would certainly be an interesting topic for future research.

Summary

Witt [1,2] expected to find an inverse relationship between mean state loss ratios and the standard deviations of the associated annual loss ratios by state for automobile insurance if underwriting risk were properly recognized in developing rates. However, he found a positive rather than an inverse relationship between these variables. Hedges argued that this observed positive relationship could have resulted from a mathematical relationship between the average loss ratios and standard deviations and/or from certain institutional factors which affect loss ratios and insurance rates differently in the various states. Venezian tried to show that there was some logical support for Hedge's mathematical hypothesis. He demonstrated that the Central Limit Theorem itself was not strong enough to imply that the sample mean and standard deviation were asymptotically uncorrelated. However, we showed that a double Central Limit Theorem was in effect in Witt's earlier empirical work, and this effect is sufficient to imply that the correlation between the sample means and standard deviations should be close to zero. This was the basic thrust of a heuristic argument by Witt in his reply to Hedges. A precise mathematical justification for this argument was developed in this paper. Moreover, it was shown that under certain conditions treating premiums as a random variable did not qualitatively change the results as Venezian thought it might. As Venezian noted, perhaps other measures of underwriting risk might be used to better explain such results. Several such measures were presented here. Perhaps, this discussion will stimulate

some additional research on the relationship between underwriting risk and return. The inclusion of investment return in a more comprehensive testable hypothesis might help to resolve the underwriting risk and return paradox.

REFERENCES

1. Witt, Robert C., The Automobile Insurance Rate Regulatory System in Illinois: A Comparative Study (Illinois Insurance Laws Study Commission, 1977).

2. Witt, Robert C., "The Competitive Rate Regulatory System in Illinois: A Comparative Study," CPCU Journal, Vol. XXXI, No. 3 (September 1978), pp. 151-162.

3. Hedges, Bob A., "On Positive Correlation Between Means and Standard Deviations of Claims Ratios: Commentary," The Journal of Risk and Insurance, Vol. XLVIII, No. 4, pp. 649-652.

4. Witt, Robert C., "Underwriting Risk and Return: Some Additional Comments," The Journal of Risk and Insurance, Vol. XLVIII, No. 4 (December 1981), pp. 653-661.

5. Kendall, Maurice G., and Alan Stuart, The Advanced Theory of Statistics, Vol. 1, Distribution Theory, Fourth Edition (New York: Macmillan Publishing Company, Inc., 1977), pp. 243-246.

6. Box, G. E. P., and G. Jenkins, Time Series Analysis, Forecasting and Control (San Francisco: Holden Day, Inc., 1970).

7. Kemperman, J. H. B., "The Oscillating Random Walk," Stochastic Processes and Their Application, Vol. II (1974), pp. 1-29.

8. Keilson, J., and J. A. Wellner, "Oscillating Brownian Motion," Journal of Applied Probability, Vol. XV (1978), pp. 300-310.

9. Aggarwal, M. L., "Records and Oscillating Random Walks," Calcutta Statistical Association Bulletin, Vol. XXV (1976). pp. 55-64.

10. Witt, Robert C., and Harry Miller, "Rate Regulation, Competition, and Underwriting Risk in Automobile Insurance Markets," CPCU Journal, Vol. XXXIV, No. 4 (December 1981), pp. 202-220.

ARL-TP-82-8
22 March 1982

DENSITY ESTIMATES OF SURFACE AND BOTTOM REVERBERATION

by

Gary R. Wilson
Dennis R. Powell

APPLIED RESEARCH LABORATORIES
THE UNIVERSITY OF TEXAS AT AUSTIN
AUSTIN, TEXAS 78712

Prepared for:

OFFICE OF NAVAL RESEARCH
N00014-81-K-0145

This paper has been submitted by its authors to The Journal of the Acoustical Society of America for publication.

Density estimates of surface and bottom reverberation

Gary R. Wilson and Dennis R. Powell

Applied Research Laboratories,

The University of Texas at Austin, Austin, TX 78712

(Received

Selected surface reverberation and bottom reverberation returns were used to compute estimates of the probability density function of the instantaneous reverberation. To estimate the densities, 6500 samples of surface reverberation and 3078 samples of bottom reverberation were used. The collections of samples were tested for randomness, independence, homogeneity, and normality. Both the surface and bottom reverberation were found to be non-Gaussian. Kernel techniques were applied to the collections of samples to provide univariate estimates of the densities. The densities were seen to be nearly Gaussian, but with heavier tails.

PACS numbers:

I. INTRODUCTION

It is commonly assumed in signal processing for underwater acoustic applications that the noise field is a Gaussian random process. The Gaussian assumption allows one to appeal to the large volume of signal processing theory devoted to Gaussian noise fields and often allows tractable analytical solutions of the signal processing algorithms. In many instances the assumption of a Gaussian noise field is reasonable. However, it has been shown that non-Gaussian noise fields can be encountered in situations of interest in underwater acoustics.[1-3] It is also known that significant performance degradations can occur when a processor optimized for the Gaussian noise field is operating in a non-Gaussian noise field.[4] To optimize these processors for a non-Gaussian noise field, it is usually not enough to know that the noise field is non-Gaussian; an estimate of the density function of the noise field is generally required, at least for the parametric processing techniques. For example, detection algorithms that employ a likelihood ratio require some estimate of the signal and noise densities.

Previous work involving a statistical analysis of underwater acoustic fields has generally been limited to testing for normality and has not provided an estimate of the density.[2,3,5,6] The present work extends this type of analysis to include univariate density estimates of selected non-Gaussian surface and bottom reverberation returns. These returns were formed into ensembles and each ensemble was determined to be independent and identically distributed. The ensembles were

determined to be non-Gaussian by the application of several tests for univariate normality.

II. EXPERIMENTAL DATA

The surface reverberation data were collected at Lake Travis Test Station (LTTS) of Applied Research Laboratories, The University of Texas at Austin (ARL:UT).[7] The data consist of 500 returns of a pulsed cw signal scattered from the wind-roughened lake surface at a 10.5^{o} grazing angle. The transmitted frequency was 80 kHz, the pulse length was 100 μsec, and the horizontal transmit beamwidth was approximately 15^{o}. Winds were from the north at approximately 35-50 km/h and had a fetch of 2.4 km. Wave heights were estimated visually to be about 0.3-0.6 m (trough to crest).

The bottom reverberation data were collected from a bottom that was very uniform and consisted of loose silty sand mixed with shell. The sonar had a horizontal transmit beamwidth of approximately 3^{o} and operated at a frequency of 18 kHz and a pulse length of 100 μsec. These data were collected at a grazing angle of 35^{o}. Fifty-seven returns were used to form the bottom reverberation ensembles. The side scan sonar was moving in a straight line during data collection, resulting in returns from different parts of the bottom.

All reverberation data were sampled digitally in time by a 12-bit analog-to-digital (A/D) converter representing a range of $\pm$10 V. The samples from each return occurring at the same time after transmission were associated together into ensembles. Thirteen ensembles of surface

reverberation were formed. Each ensemble was separated in time by 250 μsec. Fifty-four ensembles of bottom reverberation were formed, separated in time by 110 μsec.

III. ENSEMBLE VALIDATION AND TESTING

Each of the ensembles was tested for randomness, independence, homogeneity, and normality. The runs up and down test for randomness, the Kolmogorov-Smirnov two-sample test for homogeneity, the Wilcoxon rank sum test for homogeneity, Pearson's test of skewness for normality, and D'Agostino's test for normality were applied to the surface reverberation data. In addition, the Kendall rank correlation tests for randomness and independence were applied to the bottom reverberation data. Descriptions of these tests can be found elsewhere.[3,8-11]

The results of these tests are shown in Fig. 1 for the bottom reverberation and in Fig. 2 for the surface reverberation. The results of each test are displayed as the probability of a Type I error for each ensemble. The probability of a Type I error is the probability of making a mistake by announcing the alternative hypothesis when the null hypothesis is true (e.g., announcing a non-Gaussian ensemble when it is actually Gaussian). A small value of the probability of a Type I error indicates the null hypothesis of the test can be rejected with some confidence.

The probability of a Type I error is itself a random variable and is uniformly distributed on the interval $[0,1]$ if the null hypothesis is true, i.e., if α denotes the probability of a Type I error then

$$\text{Prob}[\alpha \geq P | H_0] = P \quad , \quad 0 \leq P \leq 1 \quad .$$

When many independent ensembles of the same random process are tested, their outcomes can be combined to give an overall probability that the particular hypothesis being tested is valid for the entire set of ensembles. If this probability is low (say, less than 0.05), then the null hypothesis of the test can be rejected with some confidence for the entire set of ensembles. This combined probability is indicated in the figures as P_c.

The method for combining the probabilities is Edgington's normal curve method.[12,13] The mean of the probabilities is computed and, if the probabilities are uniformly distributed on the interval $[0,1]$, then according to sampling theory the sample mean will be normally distributed with a mean of 0.5 and a standard deviation of $1/12n$, where n is the number of probabilities combined. Normal curve tables can then be referenced to determine the probability that the sample mean represents the mean of uniform $[0,1]$ random variables. For some tests, for example the runs test, the exact probability of a Type I error cannot always be computed, but only the upper and lower bounds on the probability. Thus only an upper and lower bound on P_c can be computed. For these tests, the bounds on the probability of a Type I error are indicated by two lines on the plots in Figs. 1 and 2, and the corresponding upper and lower limits of P_c are given. In some cases the upper and lower limits are the same.

As can be seen, the ensembles of both surface and bottom reverberation can be accepted as consisting of independent, identically

distributed random variables. The values of P_c for the tests for randomness, independence, and homogeneity are greater than such a low value as 0.05, indicating that the null hypotheses of these tests cannot be rejected with any degree of confidence.

However, the results of Pearson's test of skewness and D'Agostino's test for normality indicate that the hypothesis of normality can be rejected for both the surface and bottom reverberation. Although the value of P_c for Pearson's test of skewness for the surface reverberation (Fig. 2) does not indicate that the Gaussian assumption can be rejected, the results of D'Agostino's test give a confident rejection of normality. Since D'Agostino's test is more sensitive to the kurtosis of the data, these results indicate a rejection of normality primarily due to a significant kurtosis. Although not shown here, Pearson's test of kurtosis was also performed on the surface reverberation ensembles and the results also indicated a confident rejection of normality. The normality of the bottom reverberation data, Fig. 1, can be marginally rejected by Pearson's test of skewness and confidently rejected by D'Agostino's test. As shown later, both the surface and bottom reverberation had larger than normal kurtosis, indicating heavier tails than a Gaussian density.

Although the ensembles as a whole tested to be non-Gaussian, it is not clear that all the ensembles of each type of reverberation were produced from the same underlying non-Gaussian distribution. For this reason, a k sample homogeneity test[14] was applied to the 13 surface reverberation ensembles and the 54 bottom reverberation ensembles. The

results of this test indicated that the hypothesis of homogeneity could not be rejected for the 13 ensembles of surface reverberation, nor for the 54 ensembles of bottom reverberation. The probability of Type I error was 0.9536 and 0.8054, respectively. This result also is a justification for combining ensembles together during density estimation, which will be discussed in the next section.

IV. DENSITY ESTIMATION

A. Technique of density estimation

The nonparametric estimation of the probability density for a data set has been dealt with by many authors, notably Rosenblatt[15] and Parzen.[16] The two general methods used in density estimation are series approximation using orthogonal functions and series of kernel functions. The latter method is used here. If $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables, then the kernel density estimate is

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(n)} K\left(\frac{x - X_i}{h(n)}\right) \quad ,$$

where K(x) satisfies certain regularity conditions (cf. Parzen). The function h(n) defines a sequence of kernel widths.

In using estimates of this form, a choice must be made for K(x) and h(n). Many choices for the kernel function were considered, including rectangular, quadratic, and exponential forms. The kernel function used is due to Silverman:[17]

$$K(x) = \begin{cases} 1/4\ x^4 - 1/2\ |x|^3 + 1/2 & \text{if } |x| \leq 1 \\ 1/4\ |x|(2-|x|)^3 & \text{if } 1 < |x| \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

This quartic was selected for its vanishing support outside the interval $[-2,2]$.

There are two classes of choices for h(n): constant widths or variable widths. Kernel estimates have been shown to be consistent for both classes of h(n), implying that either choice will suffice for very large sample sets. If the kernel width is constant, then asymptotic results prove that the optimal kernel width is proportional to $n^{-1/5}$. Kernel estimates using variable widths have been investigated by Penrod and Machell et al., as well as others.[18,19] In this study, both variable and constant width kernel density estimates were attempted. For reasons given below, estimates based on constant width kernels are presented. The selection of the constant was initially based on the optimal h(n) defined by Rosenblatt, although ignorance of the true underlying density necessitated empirical determination of an acceptable constant.

B. Results of density estimates

Both the fixed and variable width kernel density estimation algorithms were run on the sample data. The results indicated that the fixed width estimates provided less smooth densities, while the variable width density estimates yielded very smooth densities. The comparison

of the moments of the density estimates with the sample moments of the data demonstrated large disparities for the variable width estimates based on the k nearest neighbor technique,[18] particularly in the measurements of variance and kurtosis. The fixed width estimates provided much closer agreement for all of the moment statistics. For this reason the variable width estimates were deemed unreasonable, and only densities from fixed width estimates are given. The criteria for selecting the kernel width were based on the subjective evaluation of the "smoothness" of the density estimate along with "reasonable" agreement of the moments of the density with the sample moments of the data. For the surface reverberation data, a kernel width of 112.2 was used while a width of 70.0 was used for the bottom reverberation data. Guidelines for the selection of kernel widths for density estimates may be found elsewhere.[18]

In Fig. 3 the individual density estimates for the surface reverberation data are presented in a three-dimensional plot format. Each estimate was computed on a mesh of 101 points equally spaced in an interval containing the minimum and maximum value of the entire data set. Although some variation exists in the individual estimates, the k sample homogeneity test gives confidence that the estimates are representative of the same density. Thus it is natural to calculate the mean density estimate

$$f(x) = \frac{1}{M} \sum_{j=1}^{M} f_{n,j}(x) \quad ,$$

where M is the total number of independent density estimates and $f_{n,j}(x)$ is the density estimate for the jth ensemble. f(x) is presented as a reasonable estimate of the true probability density of the underlying acoustic process. Note that f itself is a kernel estimate and hence is a density function. Figure 4 is a plot of the mean density estimate for the surface reverberation.

The average density estimate is compared to a Gaussian density of the same mean and variance as the sample data in Fig. 4 to allow a visual comparison of this density estimate to a known (i.e., Gaussian) density. To facilitate this comparison, a log transformation of the density is performed; these results are also presented in Fig. 4. This transformation tends to highlight the difference in the tails of the density. As can be seen, the density estimate appears to be nearly Gaussian, but the log plot indicates a significant deviation in the tails. The density estimate has heavier tails than the Gaussian.

To ascertain how well the density estimate represents the data set, integral estimates of the mean, variance, skew, and kurtosis based on the density estimate were computed and compared to the sample estimates of the mean, variance, skew, and kurtosis computed from the entire data set. The results are shown in Table I. Approximate confidence intervals are given in Table II to provide an estimate of error bounds for the sample mean and variance. These confidence intervals are computed in two ways. The first assumes the data to be normal (which they are not) and the intervals are

$$\left[\bar{X} - \frac{\hat{\sigma}}{\sqrt{n}} Z\left(1 - \frac{\alpha}{2}\right) \quad , \quad \bar{X} + \frac{\hat{\sigma}}{\sqrt{n}} Z\left(1 - \frac{\alpha}{2}\right)\right] \quad ,$$

for the mean and

$$\left[\frac{n\hat{\sigma}^2}{\chi_n\left(1 - \frac{\alpha}{2}\right)} \quad , \quad \frac{n\hat{\sigma}^2}{\chi_n\left(\frac{\alpha}{2}\right)}\right]$$

for the variance.[20] Here $Z(p)$ and $\chi_n(p)$ are the pth quantiles of the standard normal distribution and the chi square distribution with n degrees of freedom, respectively. The value of α was chosen to be 0.05, resulting in 95% confidence intervals. The second confidence intervals are calculated simply as $T + 2\sqrt{var(T)}$, where T is the mean or variance statistic, and where var(T) is the sampling variance of T, given approximately by[21] $var(\bar{X}) \approx \hat{\sigma}^2/n$ and $var(\hat{\sigma}^2) \approx (\hat{\gamma}_2 - 1)\,\hat{\sigma}^4/n$. Here n is the total number of samples used to estimate these moments. The means computed for the kernel densities easily reside in their respective confidence intervals. The variances for the kernel densities do not lie in the computed confidence intervals. It was observed that the variances could be lowered and brought into the confidence intervals by decreasing the kernel width, but only at the expense of the smoothness of the density estimate. Likewise the skew and kurtosis could be brought into better agreement with the sample estimates by decreasing the kernel width. The kernel widths that were chosen provided the best agreement of the moments of the densities with the sample estimates while still maintaining a reasonably smooth density estimate. Variable

width kernel techniques were also tried, but provided even less agreement with the sample estimates of the moments.

The individual density estimates of the bottom reverberation are shown in Fig. 5. The 54 ensembles of bottom reverberation have been grouped into 9 ensembles of 342 samples each to provide more samples for the individual estimates and thus improve their accuracy. The mean density estimate for the bottom reverberation is compared to a Gaussian density in Fig. 6. This density estimate also has heavier tails than the Gaussian. It can be observed that, due to a lack of sample values in a limited range within the tails, the fixed width kernel estimates a very low value for the probability density in this range. This is particularly obvious on the log plot. Note also that there appear to be a few outliers on the extremes of the tail. It cannot be determined whether this apparent outlier group is an anomaly of the sampled process, or is in fact a feature of the density. It is felt that a larger data set would provide a more accurate estimate of the density in the tails. Table I is a comparison of the integral estimates and sample estimates of the moments of the bottom reverberation. The bottom reverberation has a larger kurtosis than the surface reverberation, resulting in heavier tails and a more significant departure from normality.

V. SUMMARY

Thirteen surface reverberation ensembles of size 500 and 54 bottom reverberation ensembles of size 57 were formed and tested for randomness,

independence, homogeneity, and normality. Both the surface reverberation and the bottom reverberation were found to be non-Gaussian. A k sample homogeneity test was applied to the data to verify that the ensembles represented the same underlying distributions. Univariate density estimates of the surface reverberation and the bottom reverberation were then computed using fixed width kernel techniques and compared to a Gaussian density. The bottom reverberation density appeared to differ more from a Gaussian density than the surface reverberation density due to heavier tails.

ACKNOWLEDGMENTS

Several helpful suggestions were provided by T. W. Sager. C. S. Penrod and F. W. Machell supplied guidance on the use of the density estimation techniques. This work was sponsored by the Office of Naval Research under Contract N00014-K-81-0145.

REFERENCES

1. T. Arase and E. M. Arase, "Deep-Sea Ambient-Noise Statistics," J. Acoust. Soc. Am. 44, 1679-1684 (1968).
2. J. E. Blue, "Signal Detection in Inhomogeneous Reverberation," Applied Research Laboratories Technical Report No. 71-21 (ARL-TR-71-21), Applied Research Laboratories, The University of Texas at Austin (1973).
3. Gary R. Wilson, "Statistical Analysis of Surface Reverberation," submitted for publication in The Journal of the Acoustical Society of America (1982).
4. Arthur D. Spaulding and David Middleton, "Optimum Reception in an Impulsive Interference Environment - Part I: Coherent Detection," IEEE Trans. Commun. IT-COM 25, 910-923 (1977).
5. Marshall E. Frazer, "Some Statistical Properties of Lake Surface Reverberation," J. Acoust. Soc. Am. 64, 858-868 (1978).
6. T. D. Plemons, J. A. Shooter, and D. Middleton, "Underwater Acoustics Scattering from Lake Surfaces. I. Theory, Experiment, and Validation of the Data," J. Acoust. Soc. Am 52, 1487-1502 (1972).
7. R. Batey and B. Korts, "Lake Travis Test Station," Applied Research Laboratories, The University of Texas at Austin, 1973.

8. Gary R. Wilson, "Covariance Functions and Related Statistical Properties of Acoustic Backscattering from a Randomly Rough Air-Water Interface," Applied Research Laboratories Technical Report No. 81-23 (ARL-TR-81-23), Applied Research Laboratories, The University of Texas at Austin (1981).

9. Gary R. Wilson, "Covariance Functions and Related Statistical Properties of Acoustic Backscattering from a Randomly Rough Air-Water Interface," Ph.D. Dissertation, The University of Texas at Austin (1981).

10. D. Middleton, "Acoustic Modeling, Simulation, and Analysis of Complex Underwater Targets, II. Statistical Evaluation of Experimental Data," Applied Research Laboratories Technical Report No. 69-22 (ARL-TR-69-22), Applied Research Laboratories, The University of Texas at Austin (1969).

11. Charles R. Baker, "Some Statistical Tests for the Analysis of Sonar Data," Department of Statistics, Report No. 8-74-3, University of North Carolina, Chapel Hill (1974).

12 Robert Rosenthal, "Combining Results of Independent Studies," Psychol. Bull. 85, 185-193 (1978).

13. Eugene S. Edgington, "A Normal Curve Method for Combining Probability Values from Independent Experiments," J. Psychol. 82, 85-89 (1972).

14. J. Kiefer, "K-Sample Analogues of the Kolmogorov-Smirnov and Cramer-V Mises Tests," Ann. Math. Stat. 30, 420-447 (1959).

15. M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," Ann. Math. Stat. 33, 1065-76 (1962).

16. E. Parzen, "On Estimation of a Probability Density Function and Mode," Ann. Math. Stat. 27, 832-37 (1956).

17. B. Silverman, "Choosing the Window Width When Estimating a Density," Biometrika 65, 1-11 (1978).

18. C. S. Penrod, F. W. Machell, and T. J. Wagner, "Empirical Finite Sample Performance of Fixed and Variable Kernel Density Estimates" (to be published in IEEE Information Theory, 1982).

19. E. F. Schuster and G. Gregory, "On the Nonconsistency of Maximum Likelihood Nonparametric Density Estimates" (to appear in "Computer Science and Statistics: 13th Symposium on the Interface," Springer-Verlag).

20. P. J. Bickel and K. A. Doksum, Mathematical Statistics (Holden-Day, San Francisco, 1977), p. 159.

21. M. G. Kendall and A. Stuart, The Advanced Theory of Statistics, Vol. I (Hafner Co., New York, 1958).

TABLE I. Comparison of the sample statistics mean, variance, skew, and kurtosis with the statistics calculated from the kernel density estimates.

Statistic	Calculation Method		Surface Reverberation		Bottom Reverberation	
	Sample	Kernel	Sample	Kernel	Sample	Kernel
mean $\hat{\mu}$	$\frac{1}{n}\sum X_i$	$\int X dF_n(X)$	-11.18	-11.14	-0.21	1.88
variance $\hat{\sigma}^2$	$\frac{1}{n}\sum (X_i-\hat{\mu})^2$	$\int (X-\hat{\mu})^2\, dF_n(X)$	105963	111513	28115	33046
skew $\hat{\gamma}_1$	$\frac{1}{n\hat{\sigma}^3}\sum(X_i-\hat{\mu})^3$	$\frac{1}{\hat{\sigma}^3}\int (X-\hat{\mu})^3\, dF_n(X)$	-0.033	-0.028	-0.069	0.032
kurtosis $\hat{\gamma}_2$	$\frac{1}{n\hat{\sigma}^4}\sum(X_i-\hat{\mu})^4$	$\frac{1}{\hat{\sigma}^4}\int (X-\hat{\mu})^4\, dF_n(X)$	3.41	3.34	4.41	4.22

TABLE II. Approximate 95% confidence interval for population statistics.

		μ	σ^2
Surface Reverberation	1	(-18.93 , -3.43)	(102412 , 109703)
	2	(-19.26 , -3.10)	(101882 , 110044)
Bottom Reverberation	1	(-6.01 , 5.59)	(26762 , 29574)
	2	(-6.25 , 5.83)	(26243 , 29987)

LIST OF FIGURES

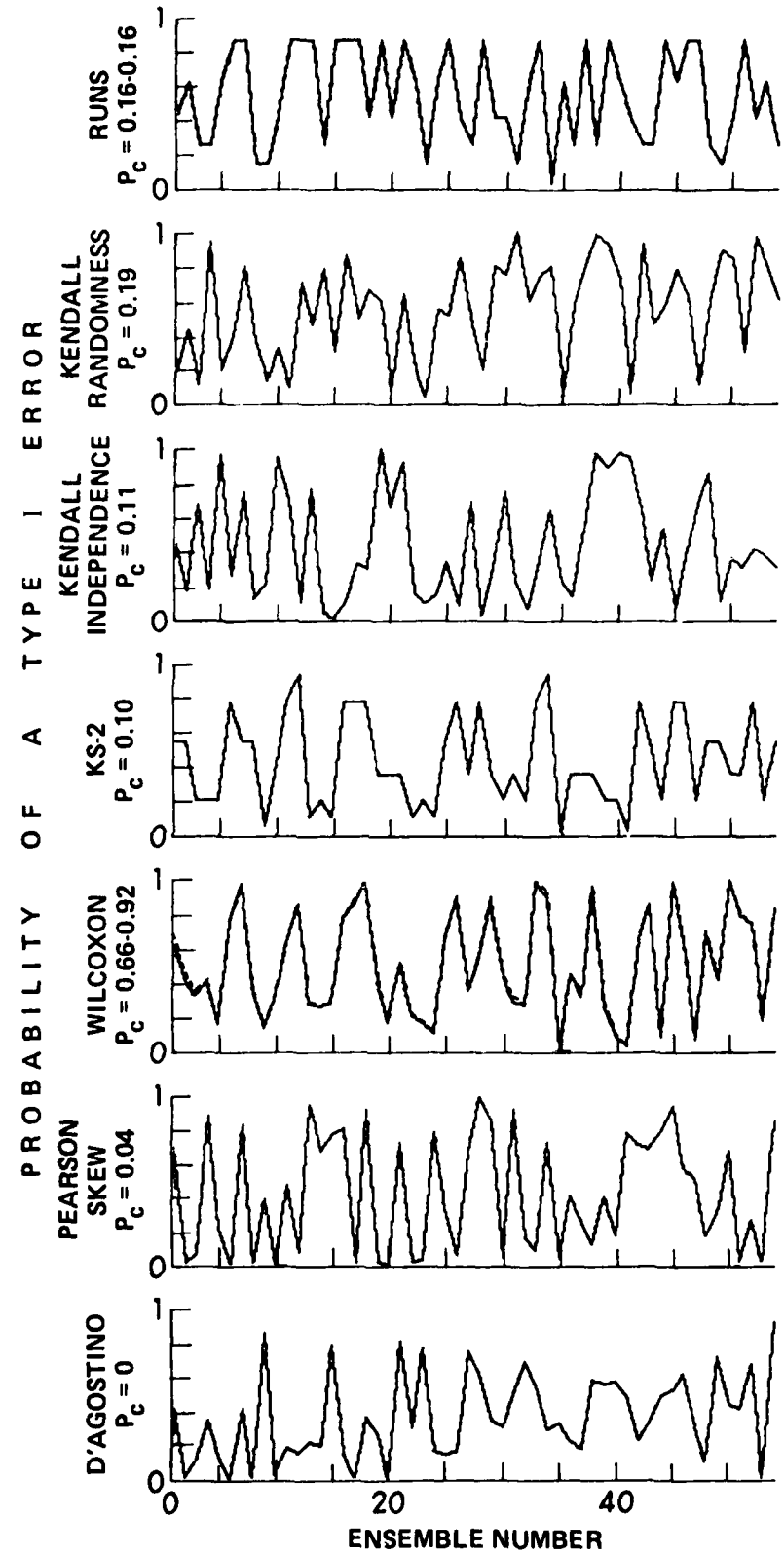


Fig. 1
Wilson

ARL:UT
AS-82-237-P
GRW-GA
3-15-82

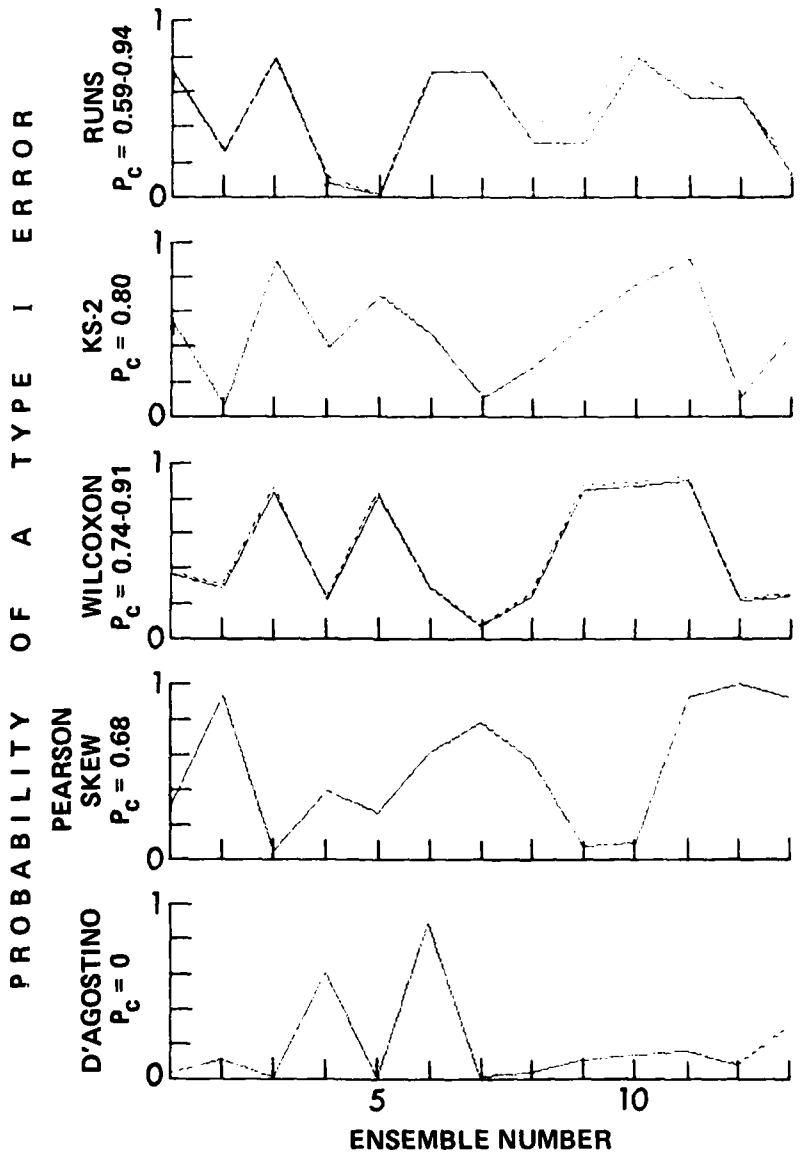


Fig. 2
Wilson

ARL:UT
AS-82-238-P
GRW-GA
3-15-82

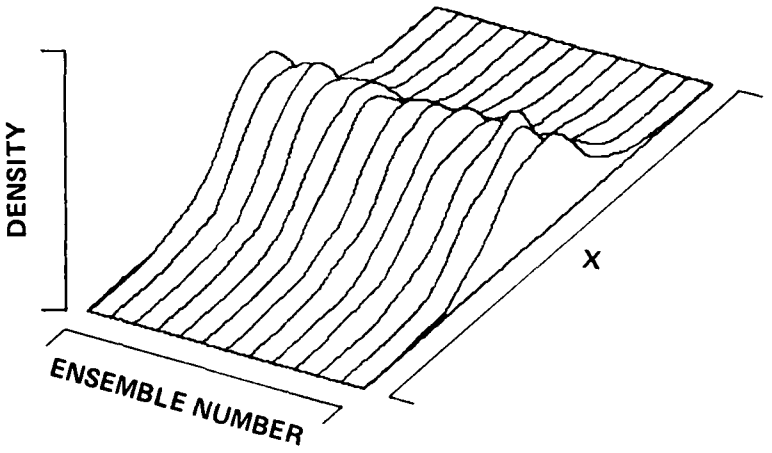
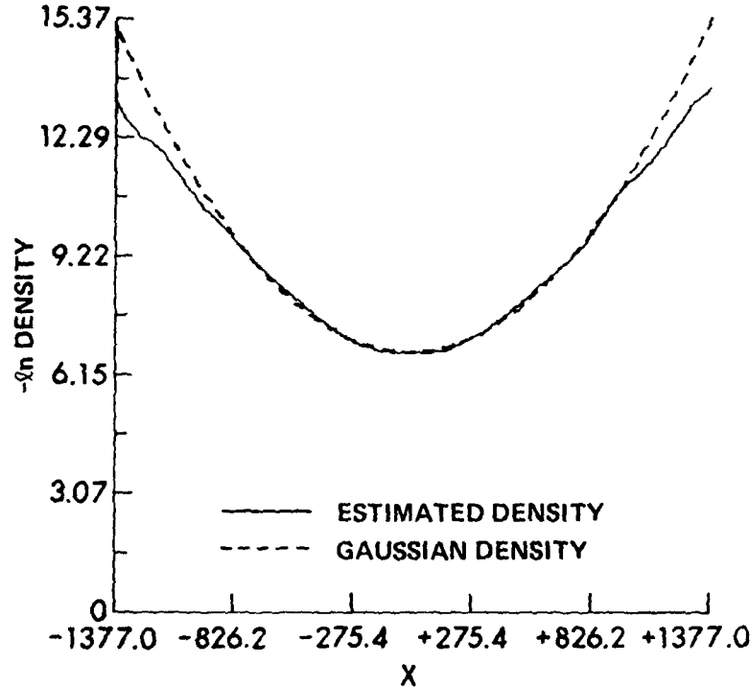


Fig. 3
Wilson

ARL:UT
AS-82-239-P
GRW-GA
3-15-82



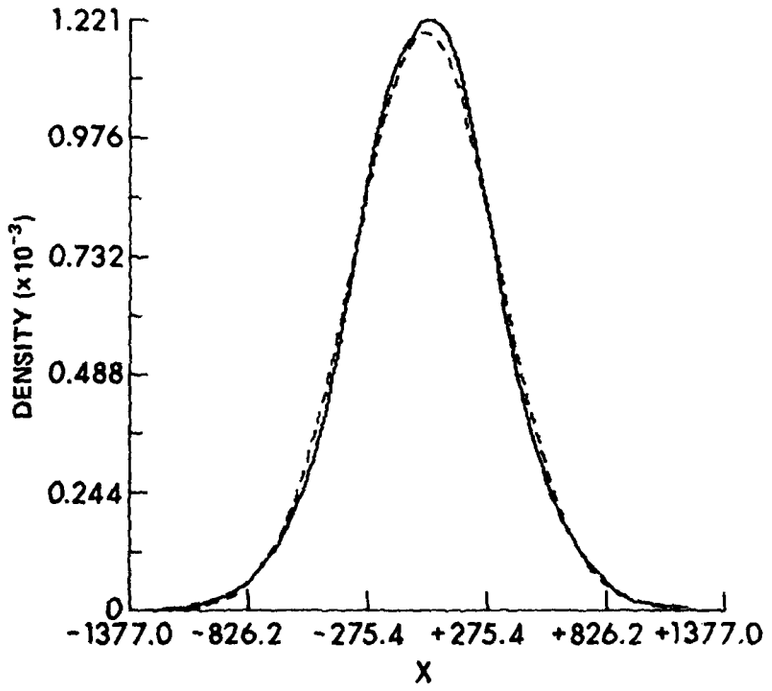


Fig. 4
Wilson

ARL:UT
AS-82-240-P
GRW-GA
3-15-82

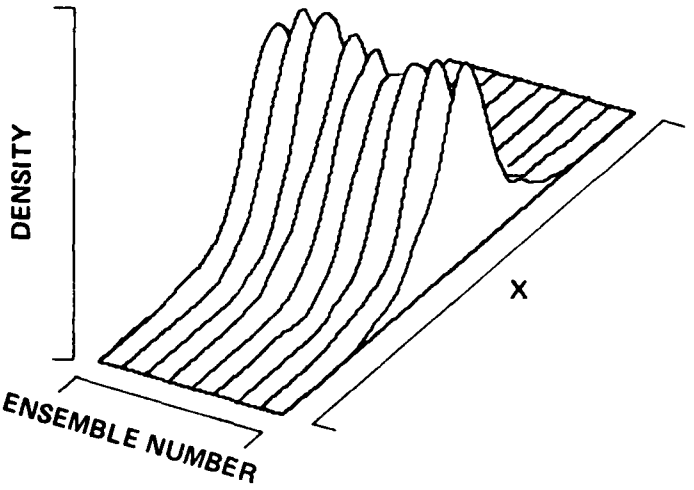
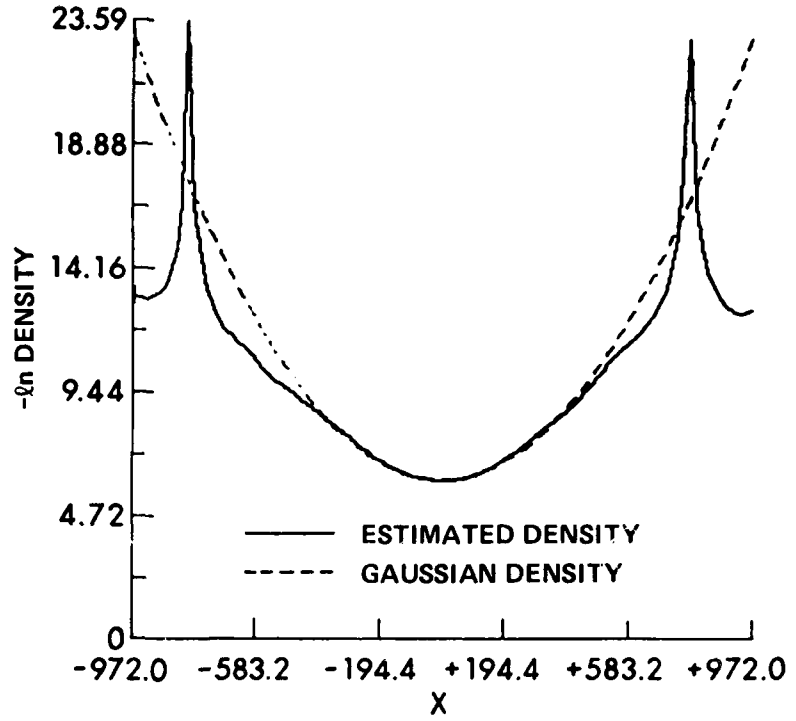


Fig. 5
Wilson

ARL:UT
AS-82-241-P
GRW-GA
3-15-82



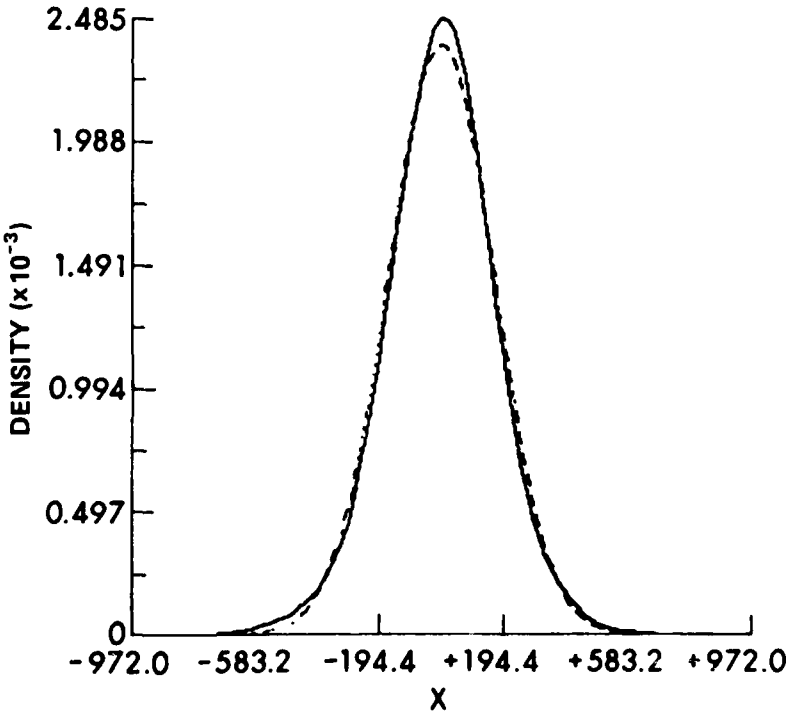


Fig. 6
Wilson

ARL:UT
AS-82-242-P
GRW-GA
3-15-82